

2017 Machine Learning with R

Naïve Bayesian Classifier

Logistic Regression

강필성

고려대학교 산업경영공학부

pilsung_kang@korea.ac.kr

목차

I

나이프 베이스 분류기

II

로지스틱 회귀분석

III

로지스틱 회귀분석 활용사례: 마케팅 반응 모델링

IV

R 실습

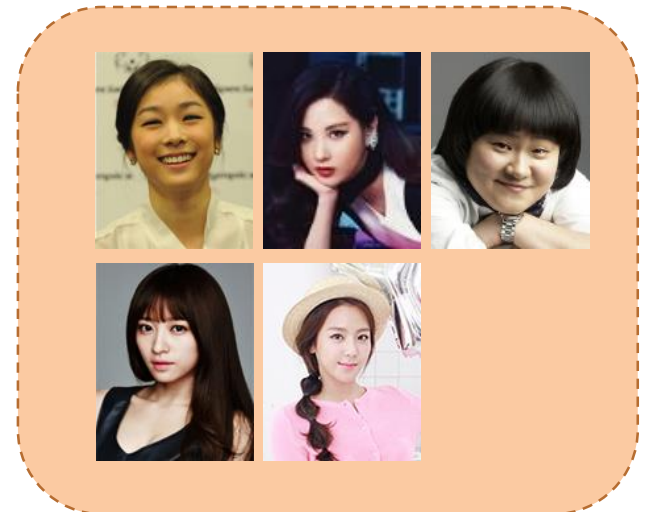
분류 문제 예시



Men

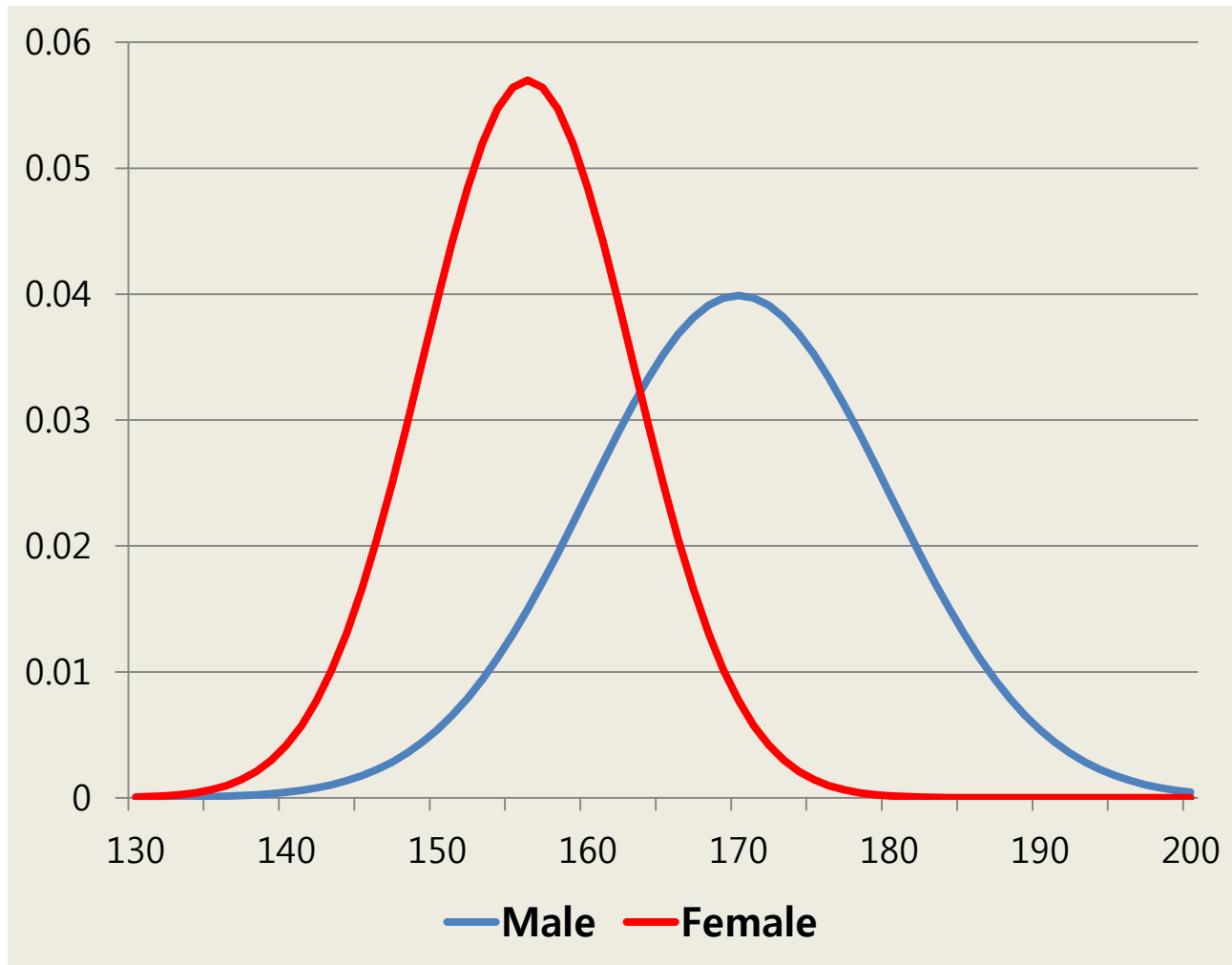
Vs.

Women



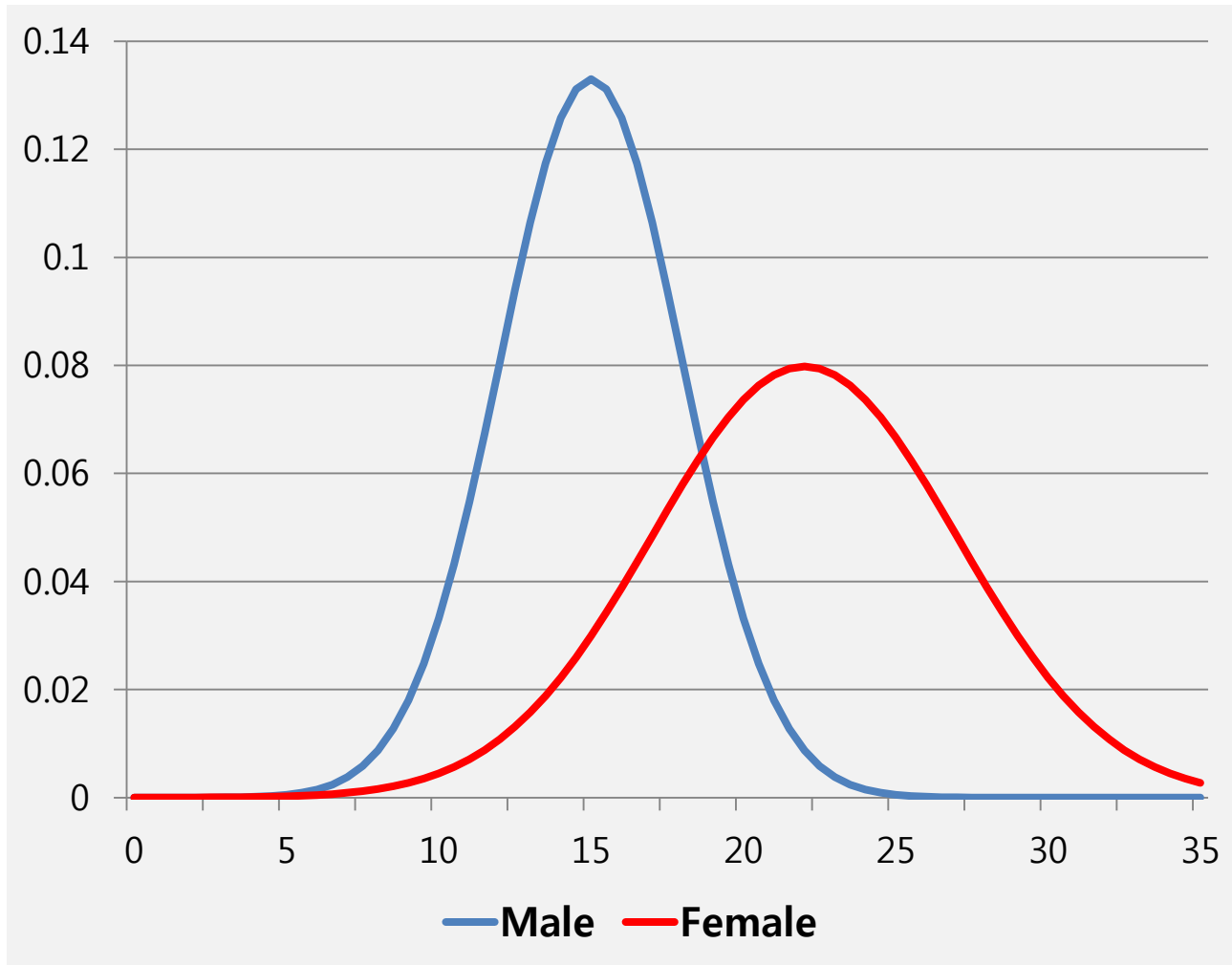
Naïve Bayesian Classification: Concept

❖ 남자와 여자의 키에 대한 사전 분포를 미리 알고 있다면...



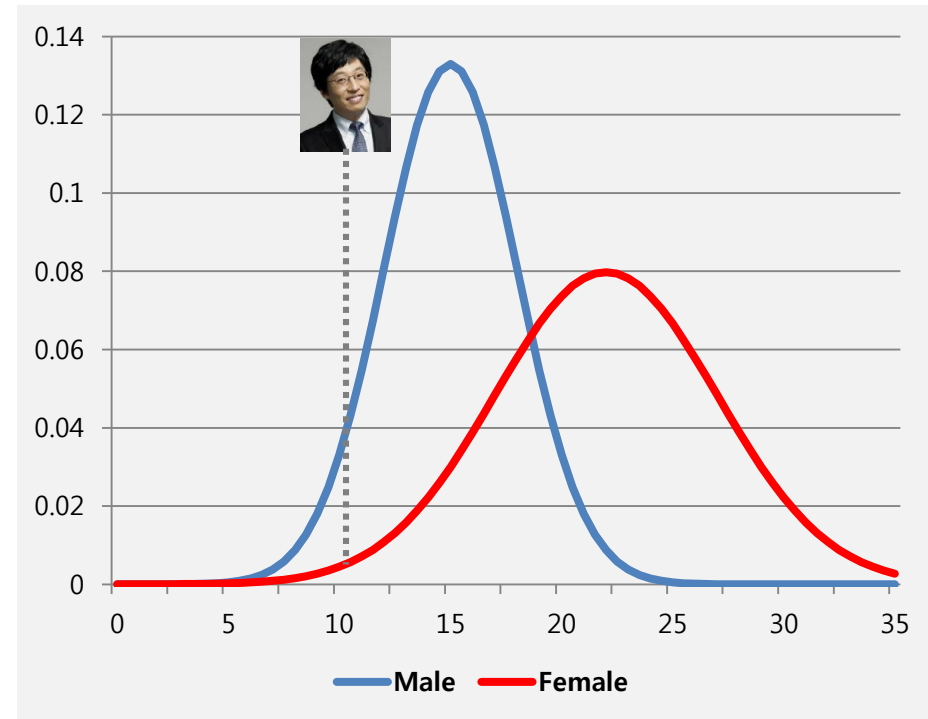
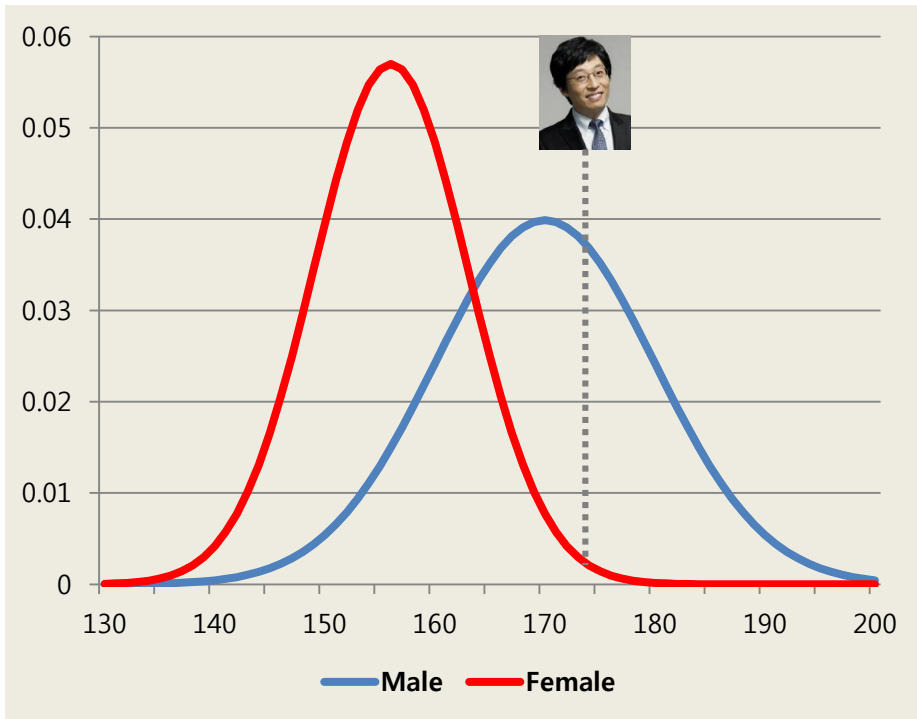
Naïve Bayesian Classification: Concept

❖ 남자와 여자의 체지방률에 대한 사전 분포를 미리 알고 있다면...



Naïve Bayesian Classification: Concept

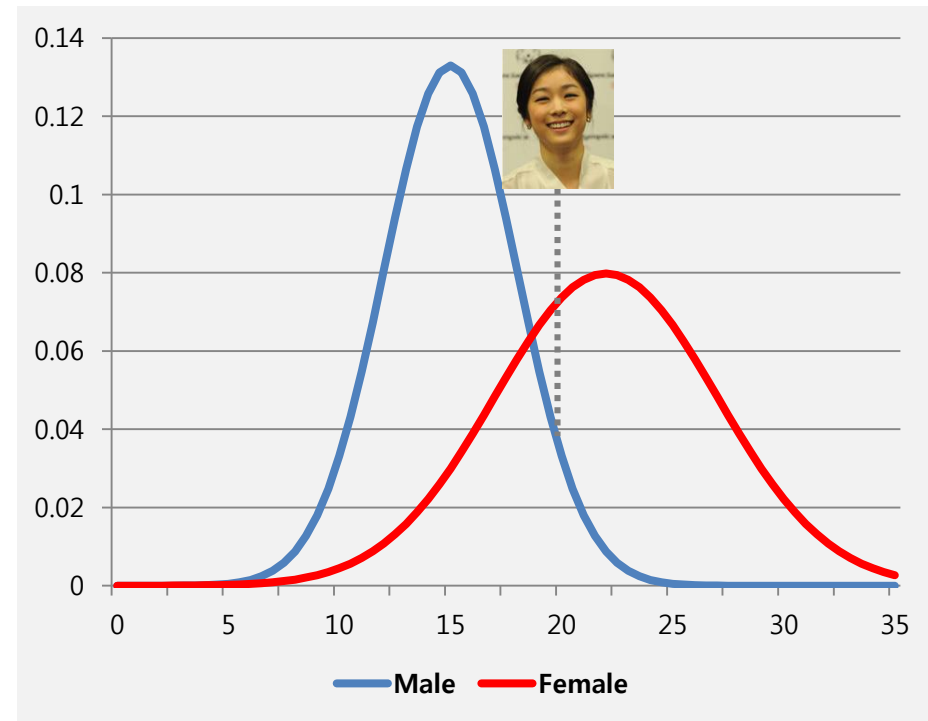
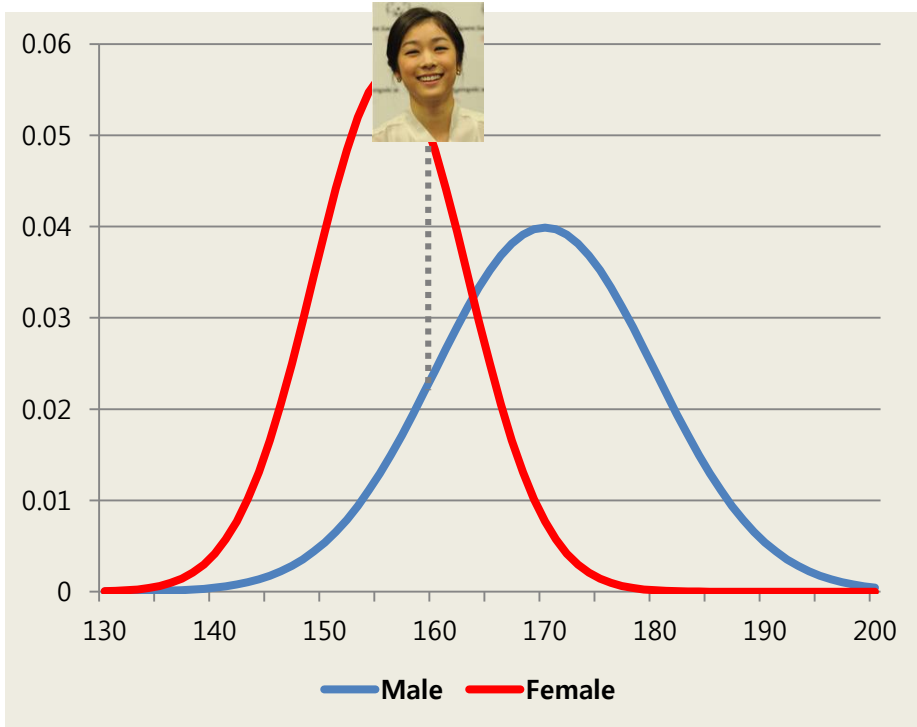
❖ 유재석은 남자일까 여자일까?



➔ 키로 보나 체지방률로 보나 남자일 가능성이 높음

Naïve Bayesian Classification: Concept

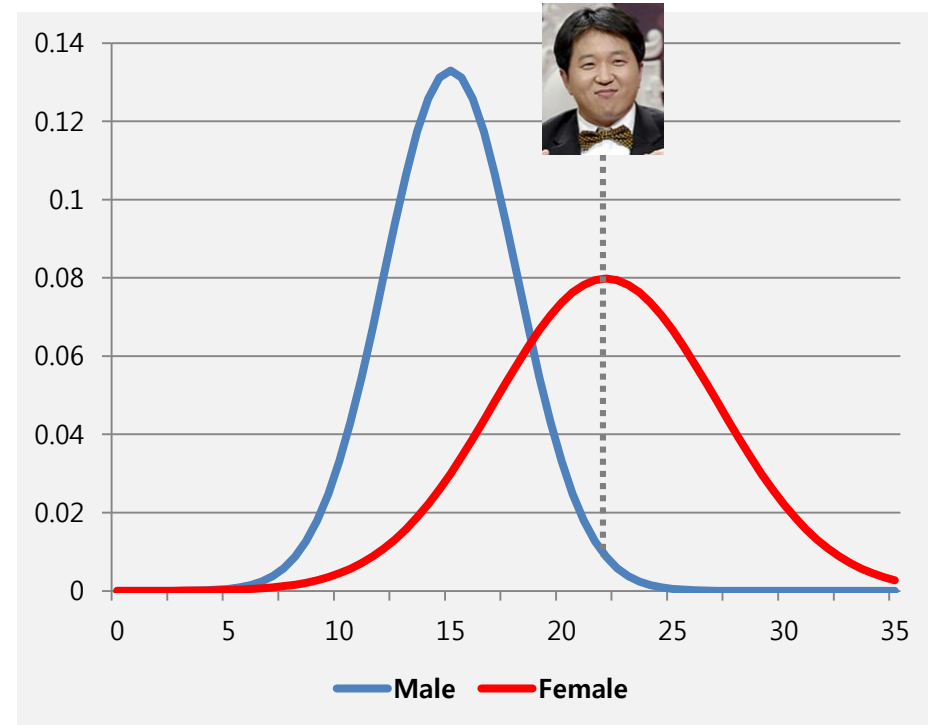
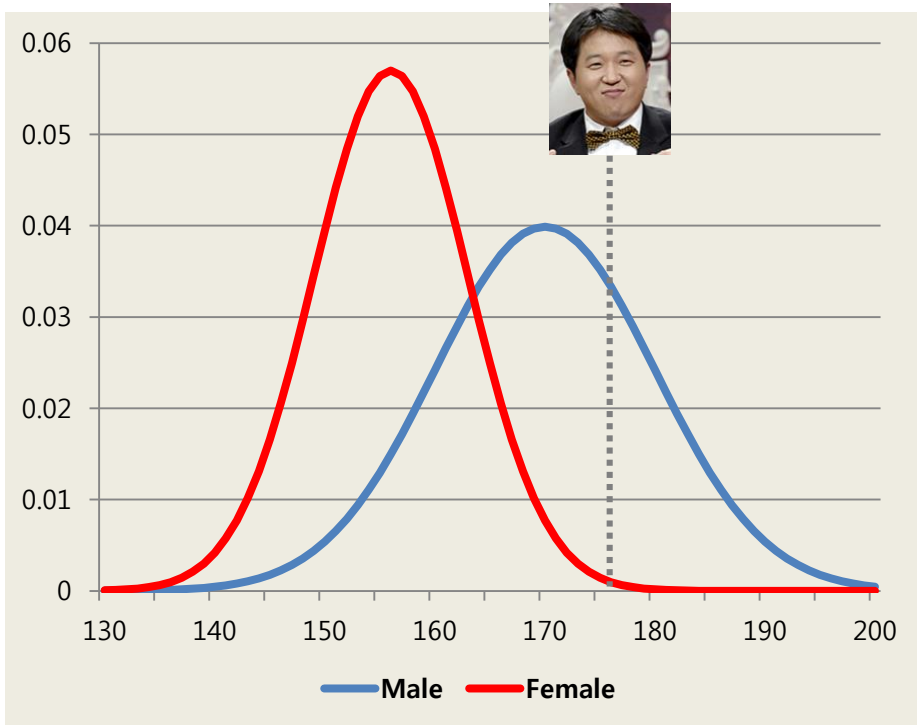
❖ 김연아는 남자일까 여자일까?



➔ 키로 보나 체지방률로 보나 여자일 확률이 높음

Naïve Bayesian Classification: Concept

❖ 그렇다면 정형돈은?



➔ 키로 보면 남자인데 체지방률로 보면 여자... 어느 범주로 분류를 해야 하지???

Naïve Bayesian Classification: Theory

❖ 베이즈 규칙(Bayes's Rule)

$$P(C_i | x) = \frac{P(C_i, x)}{P(x)} = \frac{\frac{P(x, C_i)}{P(C_i)} \cdot P(C_i)}{P(x)} = \frac{P(x | C_i)P(C_i)}{P(x)}$$

❖ 나이브 베이즈 분류기(Naïve Bayesian Classifier)

$$P(C_i | x_1, x_2, \dots, x_d) = \frac{P(x_1, x_2, \dots, x_d | C_i)P(C_i)}{P(x)} \quad \text{Baye's Rule}$$

$$= \frac{(P(x_1 | C_i) \cdot P(x_2 | C_i) \cdot \dots \cdot P(x_n | C_i))P(C_i)}{P(x)}$$

Naïve: Variables are statistically independent!

Naïve Bayesian Classification: Decision Rule

❖ 각 범주에 대한 사후 확률 계산

$$P(C_1 | x_1, x_2, \dots, x_d) = \frac{(P(x_1 | C_1) \cdot P(x_2 | C_1) \cdot \dots \cdot P(x_n | C_1))P(C_1)}{P(x)}$$

$$P(C_2 | x_1, x_2, \dots, x_d) = \frac{(P(x_1 | C_2) \cdot P(x_2 | C_2) \cdot \dots \cdot P(x_n | C_2))P(C_2)}{P(x)}$$

❖ 사후 확률값이 높은 범주로 분류

Naïve Bayesian Classification: Concept

❖ 우리가 여기서 구하고자 하는 것은 다음의 두 확률

- $P(\text{정형돈 Height, 정형돈 BFP} \mid \text{Male}) * P(\text{Male})$ vs.
- $P(\text{정형돈 Height, 정형돈 BFP} \mid \text{Female}) * P(\text{Female})$

❖ 만일 두 속성인 height 와 BFP가 통계적으로 독립이라는 가정을 할 수 있고 남자와 여자의 비율이 같다면

- $P(\text{정형돈 Height, 정형돈 BFP} \mid \text{Male}) * P(\text{Male}) =$
 $P(\text{정형돈 Height} \mid \text{Male}) * P(\text{정형돈 BFP} \mid \text{Male}) * P(\text{Male}) = 0.035 * 0.01 * 0.5 = 0.000175$
- $P(\text{정형돈 Height, 정형돈 BFP} \mid \text{Female}) P(\text{Female}) =$
 $P(\text{정형돈 Height} \mid \text{Female}) * P(\text{정형돈 BFP} \mid \text{Female}) * P(\text{Female}) = 0.001 * 0.08 * 0.5 = 0.00004$

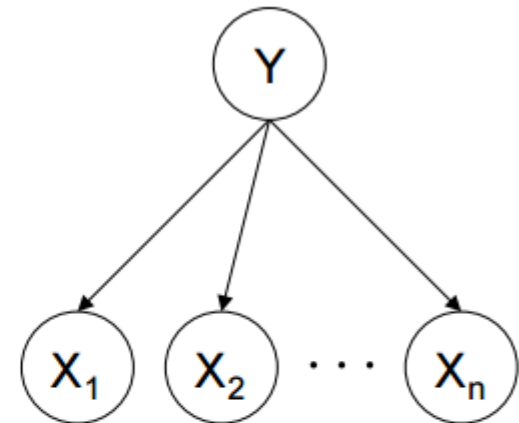
❖ $0.000175 > 0.00004$ 이므로 정형돈은 남자로 분류!

Naïve Bayesian Classification: Graphical Representation

❖ Graphical Representation of Naïve Bayesian Classifier

- **Given:**

- Prior $P(Y)$
- n conditionally independent features \mathbf{X} given the class Y
- For each X_i , we have likelihood $P(X_i|Y)$



- **Decision rule:**

$$\begin{aligned} y^* = h_{NB}(\mathbf{x}) &= \arg \max_y P(y) P(x_1, \dots, x_n | y) \\ &= \arg \max_y P(y) \prod_i P(x_i | y) \end{aligned}$$

Exact Bayesian Classifier

I

전체 레코드 중에서 테스트 개체와 정확하게 일치하는 정보를 가진 레코드들만을 선택

- 키와 체지방률이 정확히 일치하는 사람들 선택

Person	Height	BFS	Class
홍길동	178	11	M
김영희	178	11	F
김철수	178	11	M
김가네	178	11	M

Variables are not assumed to be statistically independent

$$P(C_i | x_1, x_2, \dots, x_d) = \frac{P(x_1, x_2, \dots, x_d | C_i)P(C_i)}{P(x)}$$



Exact Bayesian Classifier

2

설명변수 값이 정확히 일치하는 사람들의 주 범주 조사



Person	Height	BFS	Class
홍길동	178	11	M
김영희	178	11	F
김철수	178	11	M
김가네	178	11	M

- 3 males and 1 female.

Exact Bayesian Classifier

3

테스트 개체를 해당 범주로 분류



Person	Height	BFS	Class
홍길동	178	11	M
김영희	178	11	F
김철수	178	11	M
김가네	178	11	M

- 3 males and 1 female.
- He is classified as male.

학습 데이터에 레코드의 수가 상대적으로 적고, 설명변수의 수가 많을 경우에는 모든 설명변수의 값이 정확히 일치하는 레코드를 찾는 것이 거의 불가능함

Naïve Bayesian Classification: Procedure

I

학습 데이터 준비

- 설명변수를 정의하고 필요한 학습 데이터 수집
 - ✓ 학습 데이터 총 개체 수: 200 (남성 100명, 여성 100명)
 - ✓ 설명변수: 키(Height), 체지방률(BFS)

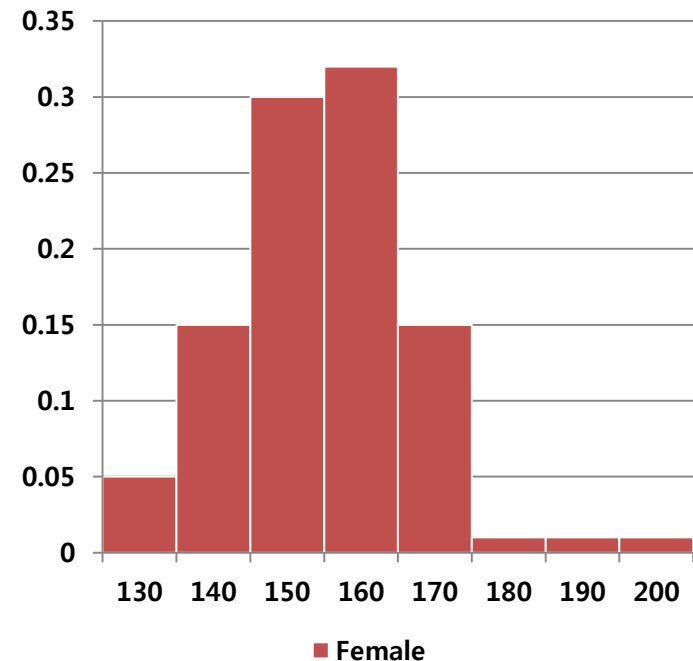
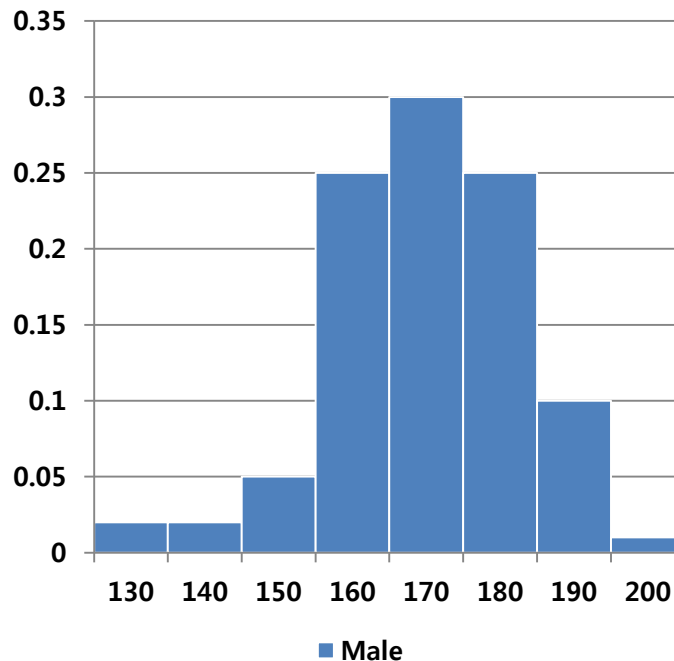
Record	Height	BFS	Class
1	187	15	M
2	165	25	F
3	174	14	M
4	156	29	F
...
N	168	12	M

Naïve Bayesian Classification: Procedure

2

범주-변수별 확률분포 추정

- 각 범주의 모든 변수에 대해 확률분포 추정: 히스토그램 사용
- 키에 대한 히스토그램

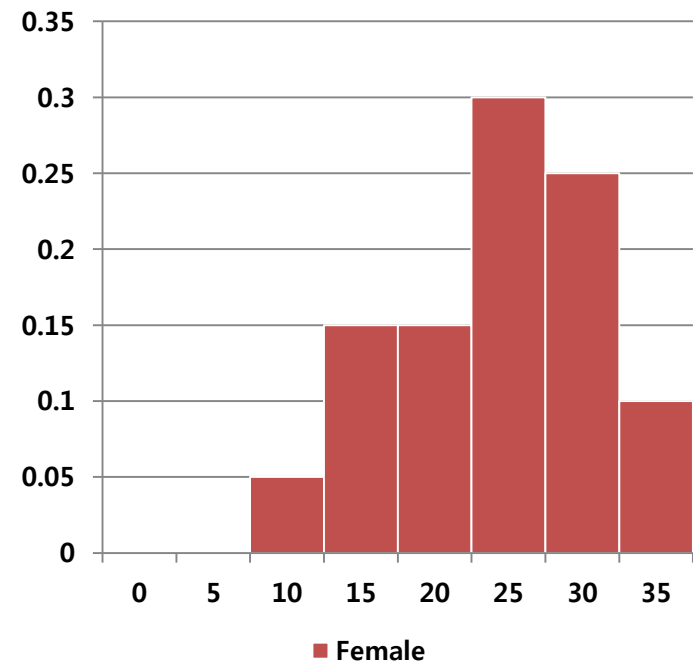
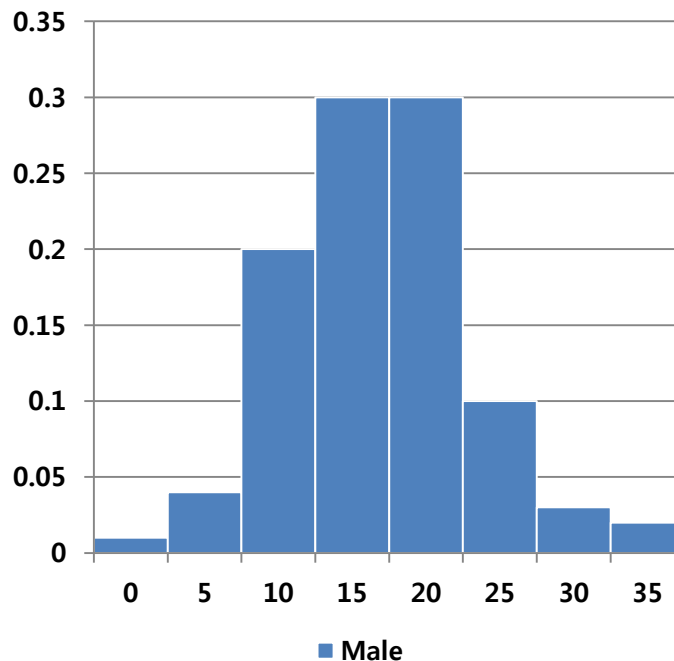


Naïve Bayesian Classification: Procedure

2

범주-변수별 확률분포 추정

- 각 범주의 모든 변수에 대해 확률분포 추정: 히스토그램 사용
- 체지방률에 대한 히스토그램

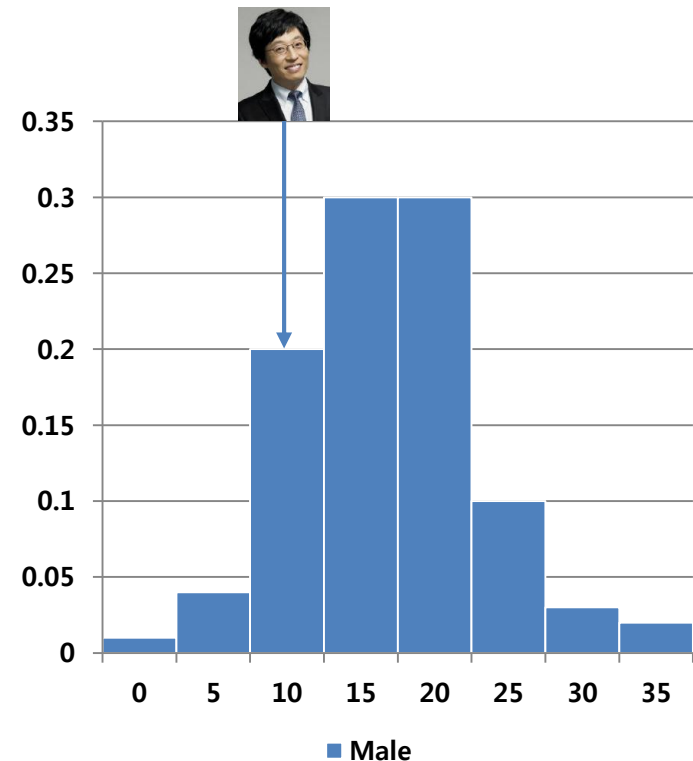
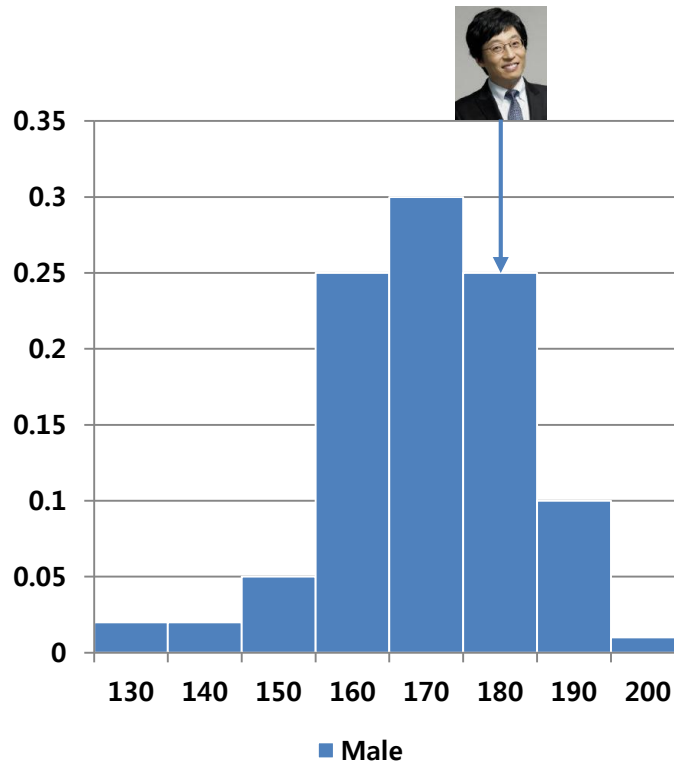


Naïve Bayesian Classification: Procedure

3

각 변수에 대한 조건부 확률 추정

- $P(\text{Height} = 178 \mid \text{Male}) = 0.25$, $P(\text{BFS} = 11 \mid \text{Male}) = 0.2$

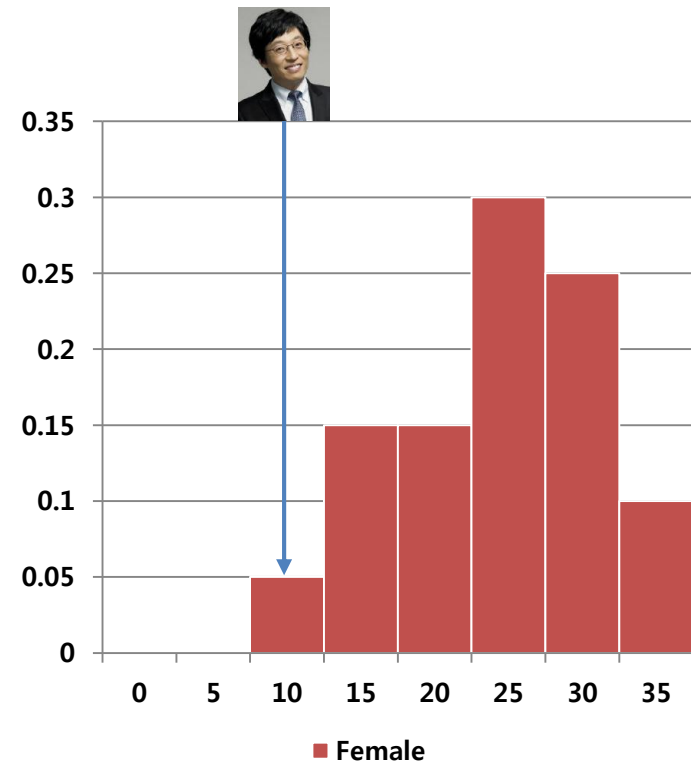
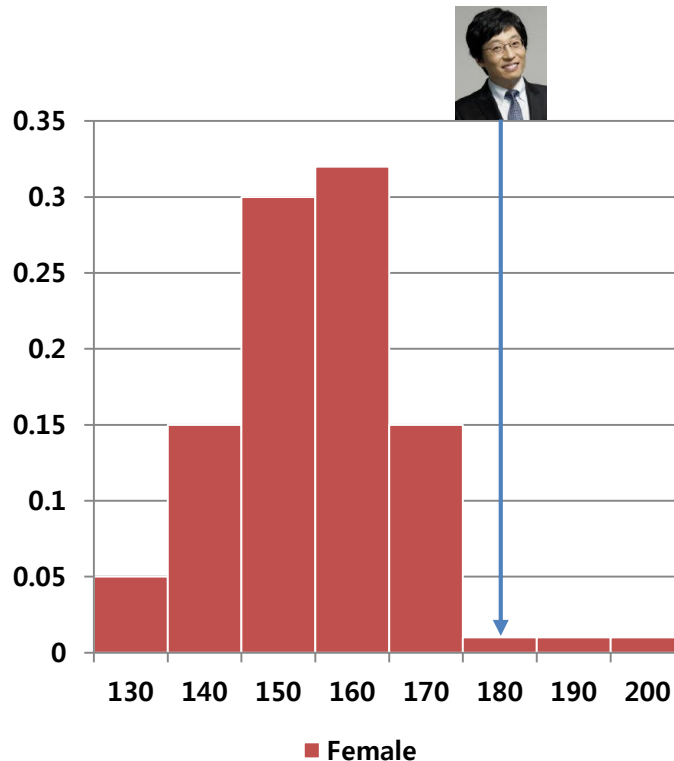


Naïve Bayesian Classification: Procedure

각 변수에 대한 조건부 확률 추정

- $P(\text{Height} = 178 \mid \text{Female}) = 0.01$, $P(\text{BFS} = 11 \mid \text{Female}) = 0.05$

3



Naïve Bayesian Classification: Procedure

각 범주에 속할 사후 확률(Posterior Probability) 추정

▪ 각 범주에 대한 사후 확률

$$✓ P(\text{Height} = 178, \text{BFS} = 11 \mid \text{Male}) * P(\text{Male})$$

$$= P(\text{Height} = 178 \mid \text{Male}) * P(\text{BFS} = 11 \mid \text{Male}) * P(\text{Male})$$

$$= 0.25 * 0.2 * 0.5 = 0.025$$

$$✓ P(\text{Height} = 178, \text{BFS} = 11 \mid \text{Female}) * P(\text{Female})$$

$$= P(\text{Height} = 178 \mid \text{Female}) * P(\text{BFS} = 11 \mid \text{Female}) * P(\text{Female})$$

$$= 0.01 * 0.05 * 0.5 = 0.00025$$

Naïve Bayesian Classification: Procedure

최종 범주 예측

- $P(\text{Height}=178, \text{BFS}=11 \mid \text{Male}) * P(\text{Male}) > P(\text{Height}=178, \text{BFS}=11 \mid \text{Female})$

$P(\text{Female}) \rightarrow$ 남성으로 분류

- 만일 학습 데이터가 400명의 남성과 100명의 여성으로 구성되어 있다면?

✓ 각 범주의 사전확률 고려: $P(\text{Male})$ & $P(\text{Female})$

✓ $P(\text{Height}=178, \text{BFS}=11 \mid \text{Male}) * P(\text{Male}) = 0.05 * 0.8 = 0.04$

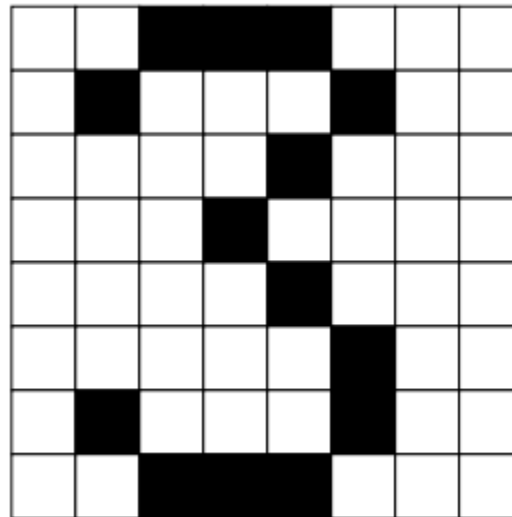
✓ $P(\text{Height}=178, \text{BFS}=11 \mid \text{Female}) * P(\text{Female}) = 0.0005 * 0.2 = 0.0001$

Naïve Bayesian Classifier Example

❖ 손으로 쓴 숫자를 판별하는 문제

- 입력 변수: 픽셀 정보
- 범주: 0부터 9까지 10개 범주

0
1
2
/
0
0



Which digit?

Naïve Bayesian Classifier Example

❖ 변수(속성 정의)

- 각 격자의 위치인 $\langle i, j \rangle$ 에 대해 값을 부여
- 단순히 1/0으로 사용할 수도 있고, 어두운 정도를 연속형 숫자로 표현할 수도 있음
- 각 이미지는 아래와 같이 벡터 형태로 변환

$$1 \rightarrow \langle F_{0,0} = 0 \ F_{0,1} = 0 \ F_{0,2} = 1 \ F_{0,3} = 1 \ F_{0,4} = 0 \ \dots F_{15,15} = 0 \rangle$$

❖ 나이브 베이저안 분류기

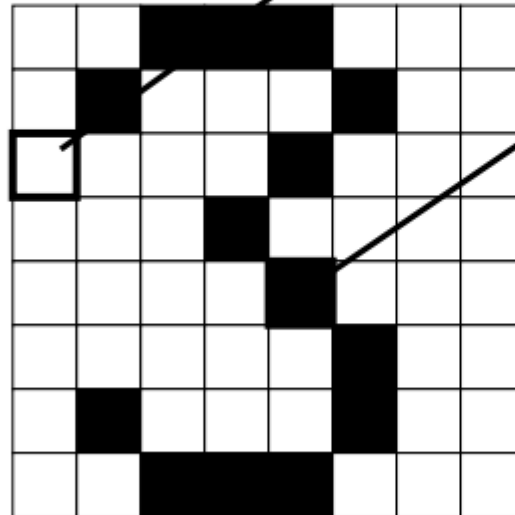
$$P(Y|F_{0,0} \dots F_{15,15}) \propto P(Y) \prod_{i,j} P(F_{i,j}|Y)$$

Naïve Bayesian Classifier Example

❖ 추정해야 되는 값은 무엇인가?

$$P(Y)$$

1	0.1
2	0.1
3	0.1
4	0.1
5	0.1
6	0.1
7	0.1
8	0.1
9	0.1
0	0.1



$$P(F_{3,1} = on|Y) \quad P(F_{5,5} = on|Y)$$

1	0.01
2	0.05
3	0.05
4	0.30
5	0.80
6	0.90
7	0.05
8	0.60
9	0.50
0	0.80

1	0.05
2	0.01
3	0.90
4	0.80
5	0.90
6	0.90
7	0.25
8	0.85
9	0.60
0	0.80

Naïve Bayesian Classifier Example

❖ 학습 절차

- 각 격자에 대해 범주의 비율을 구함

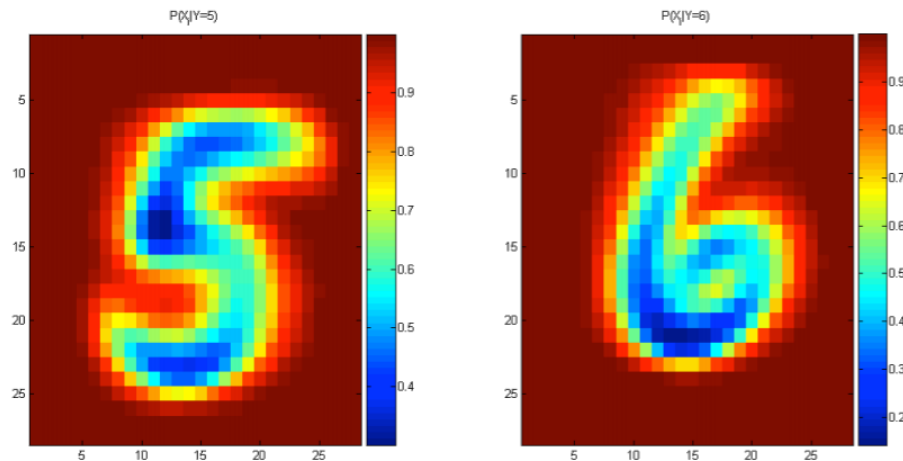
– Prior:

$$P(Y = y) = \frac{\text{Count}(Y = y)}{\sum_{y'} \text{Count}(Y = y')}$$

– Observation distribution:

$$P(X_i = x|Y = y) = \frac{\text{Count}(X_i = x, Y = y)}{\sum_{x'} \text{Count}(X_i = x', Y = y)}$$

❖ 학습된 예제



목차

I

나이프 베이즈 분류기

II

로지스틱 회귀분석

III

로지스틱 회귀분석 활용사례: 마케팅 반응 모델링

IV

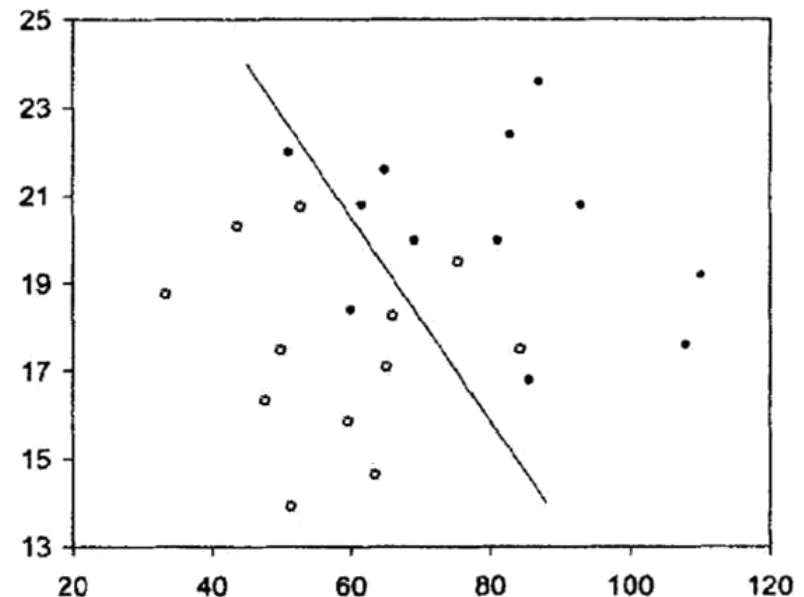
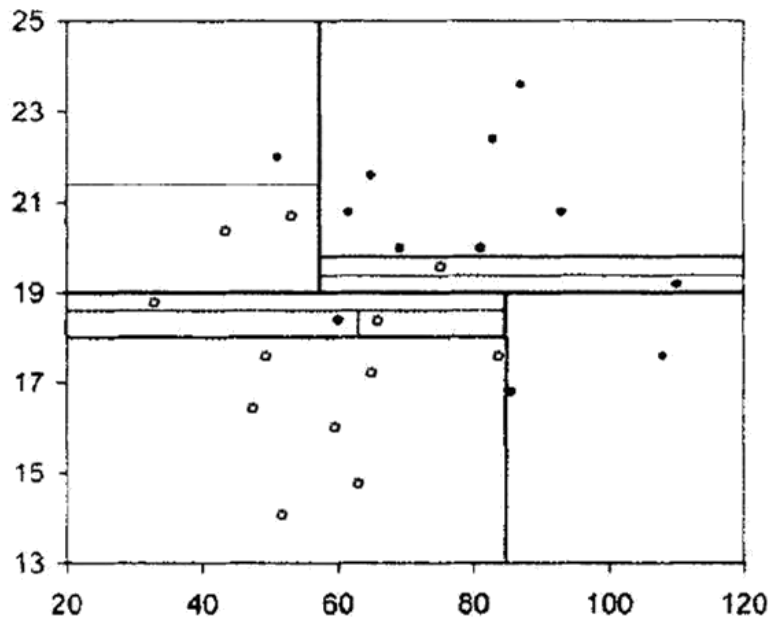
R 실습

분류 알고리즘 경계면

❖ 분류를 수행하기 위한 알고리즘은 여러 가지가 존재

- 동일한 결과를 얻기 위한 다양한 길이 존재하기 때문

“Separate the riding mower buyers(●) from non-buyers(○)”



다중선형회귀분석

❖ 목적

- 수치형 설명변수 X 와 종속변수 Y 간의 관계를 선형으로 가정하고 이를 가장 잘 표현할 수 있는 회귀 계수를 추정

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_p x_p + \varepsilon$$



다중선형회귀분석

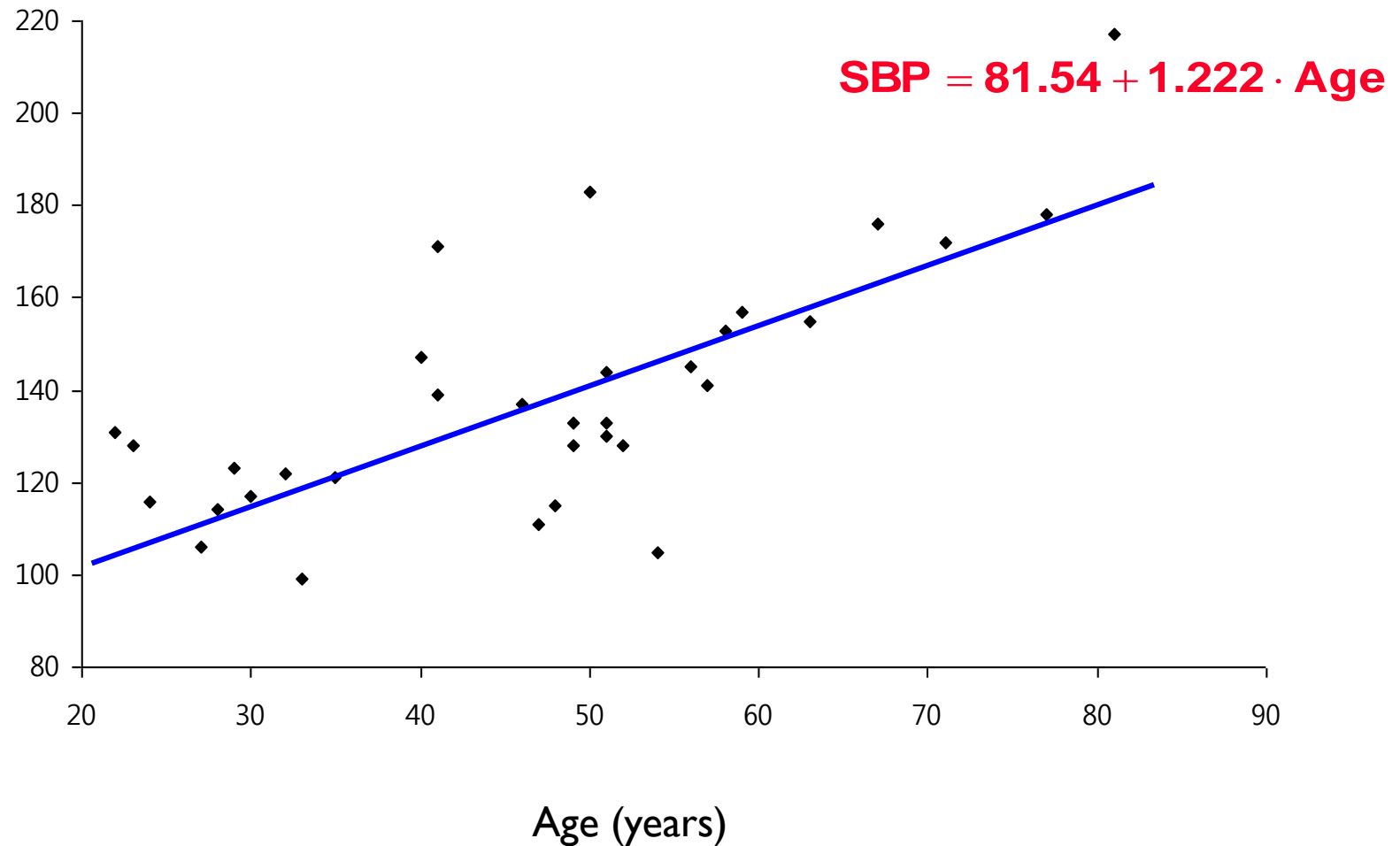
❖ 예시 I

- 33명의 성인 여성에 대한 나이와 혈압 사이의 관계

Age	SBP	Age	SBP	Age	SBP
22	131	41	139	52	128
23	128	41	171	54	105
24	116	46	137	56	145
27	106	47	111	57	141
28	114	48	115	58	153
29	123	49	133	59	157
30	117	49	128	63	155
32	122	50	183	67	176
33	99	51	130	71	172
35	121	51	133	77	178
40	147	51	144	81	217

다중선형회귀분석

SBP (mm Hg)



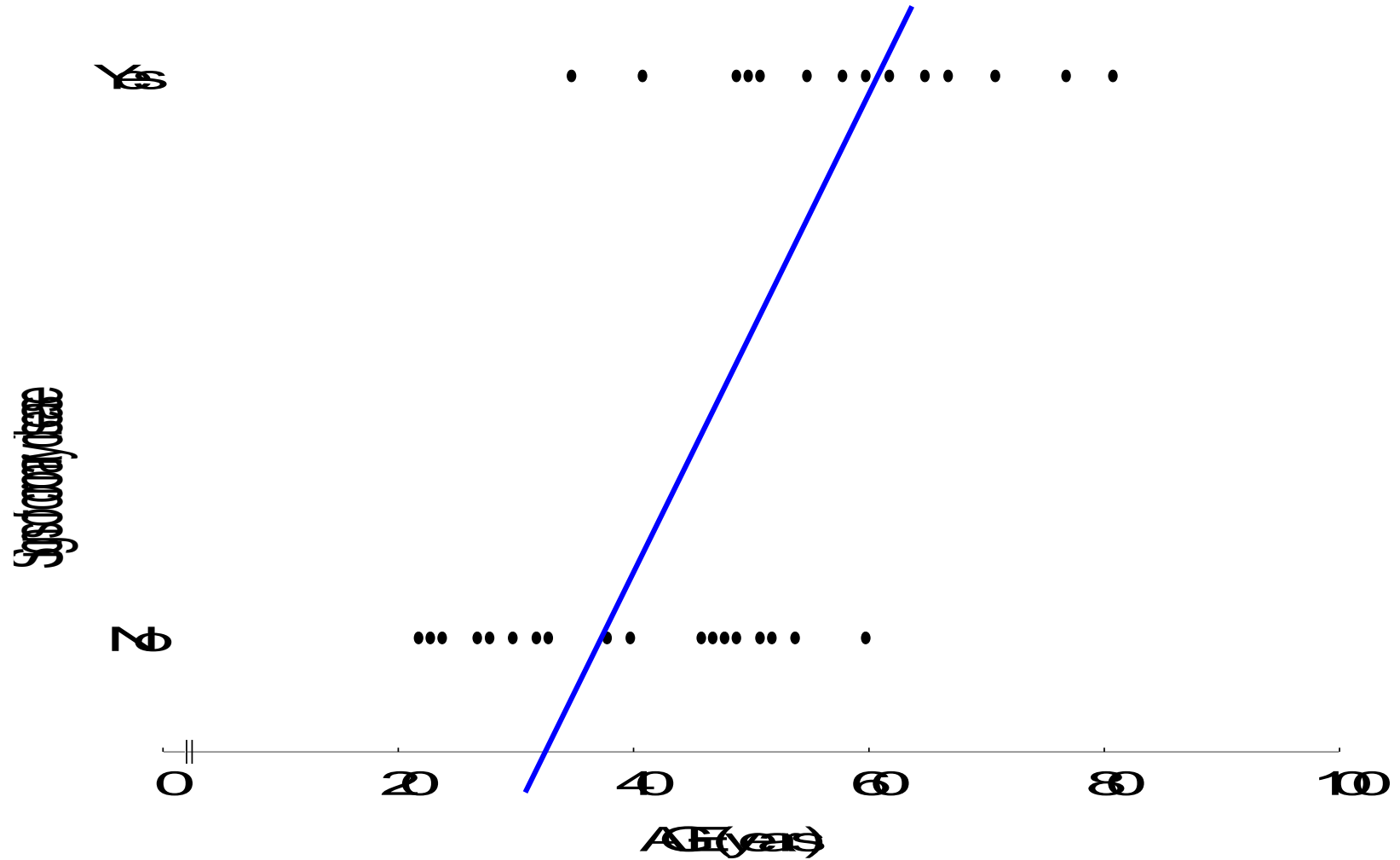
만약에...

❖ 예시 2

- 연속형 변수가 아닌 이진형(Binary) 변수인 Cancer Diagnosis를 사용한다면?

Age	CD	Age	CD	Age	CD
22	0	40	0	54	0
23	0	41	1	55	1
24	0	46	0	58	1
27	0	47	0	60	1
28	0	48	0	60	0
30	0	49	1	62	1
30	0	49	0	65	1
32	0	50	1	67	1
33	0	51	0	71	1
35	1	51	1	77	1
38	0	52	0	81	1

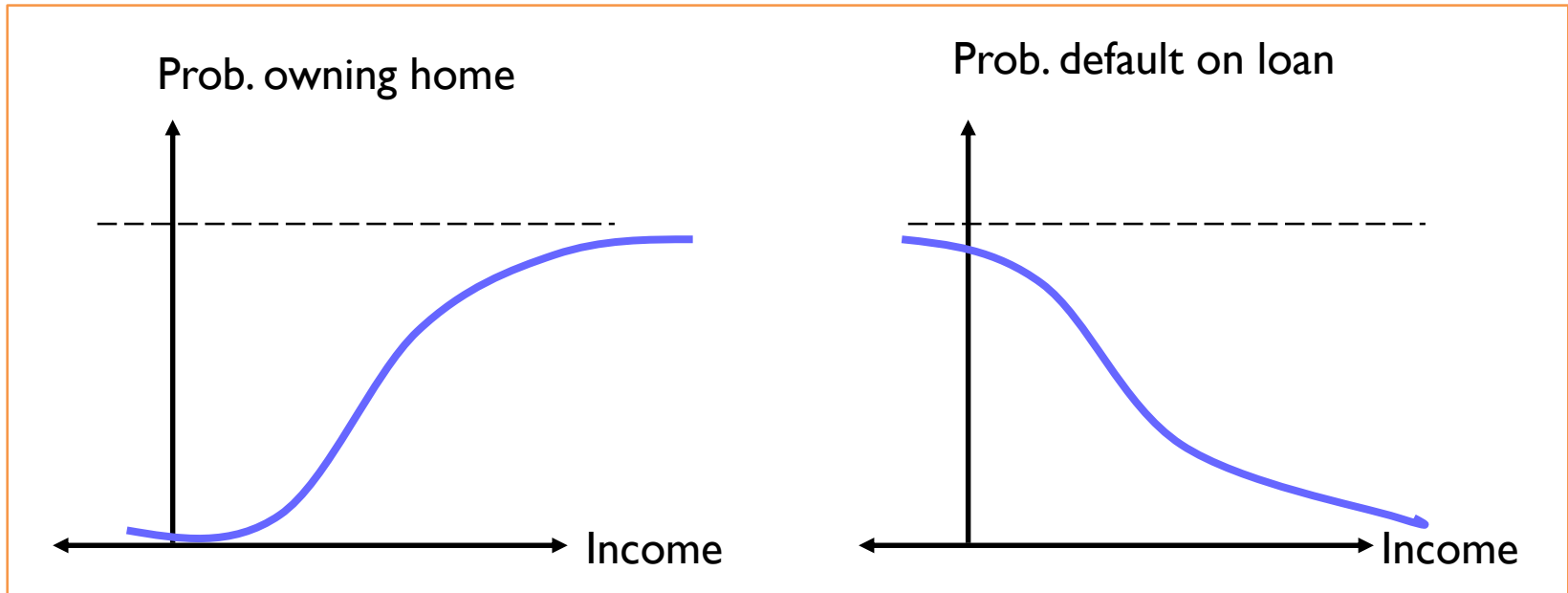
만약에...



분류 문제의 경우

❖ 실제 상황에서는

- 특정 변수에 대한 확률 값은 선형이 아닌 S-커브 형태를 따르는 경우가 많음

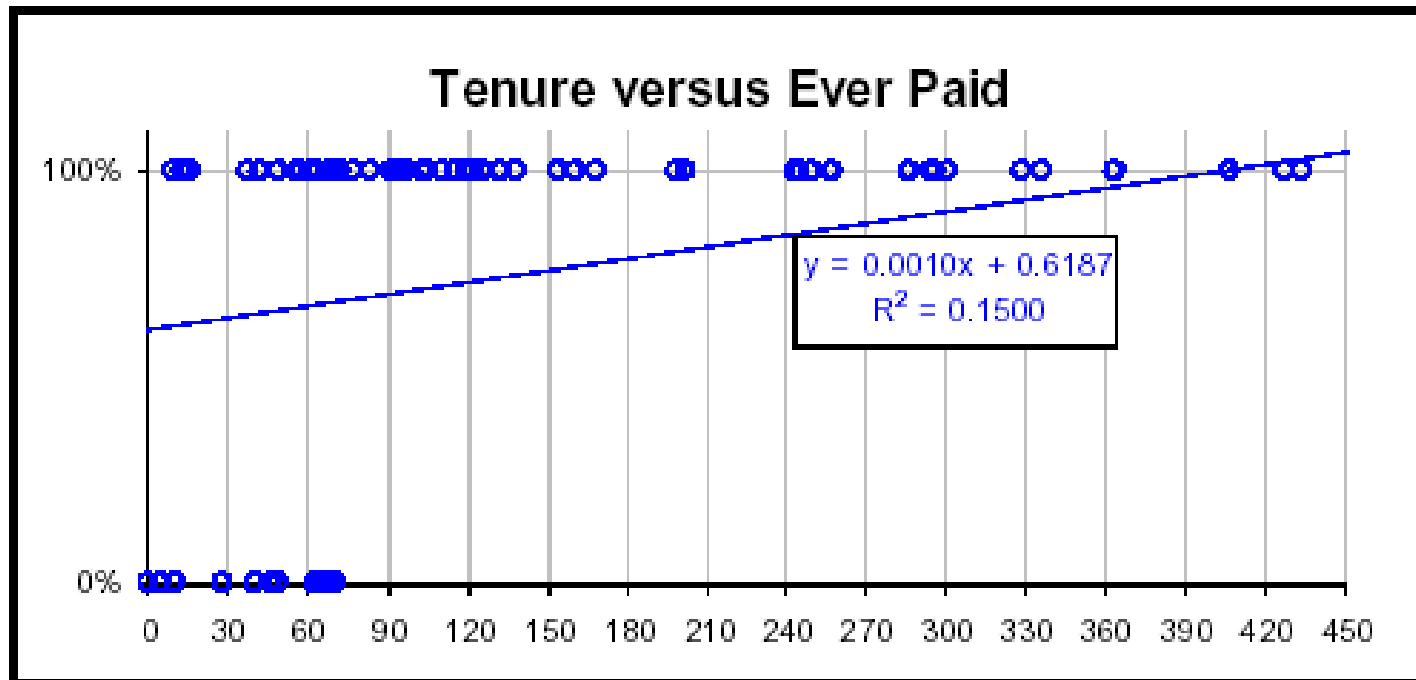


분류 문제의 경우

❖ 확률값을 선형회귀분석의 종속 변수로 사용하는 것이 타당한가?

- 선형회귀분석의 우변의 범위에 대한 제한이 없기 때문에 종속변수(좌변) 역시 범위의 제한을 받지 않음

$$P(Y = 1) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_p x_p + \varepsilon$$



로지스틱 회귀분석

❖ 목적

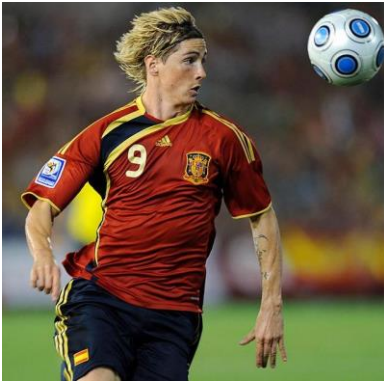
- 이진형(0/1)의 형태를 갖는 종속변수(분류문제)에 대해 회귀식의 형태로 모형을 추정하는 것

❖ 속성

- 종속변수 Y 자체를 그대로 사용하는 것이 아니라 Y 에 대한 로짓 함수(logit function)를 회귀식의 종속변수로 사용
- 로짓함수는 설명변수의 선형결합으로 표현될 수 있음
- 로짓함수의 값은 종속변수에 대한 성공 확률로 역산될 수 있으며, 이는 따라서 분류 문제에 적용할 수 있음

로지스틱 회귀분석

❖ 2010 World Cup Betting Odds



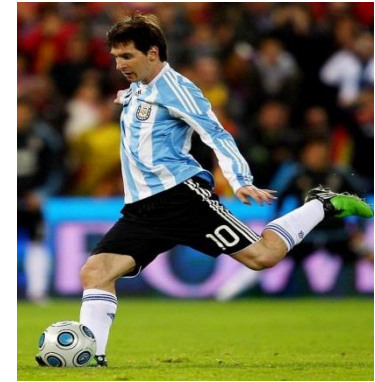
9 : 2



9 : 2



6 : 1



9 : 1



200 : 1



250 : 1



500 : 1



1000 : 1

로지스틱 회귀분석: Odds

❖ Odds

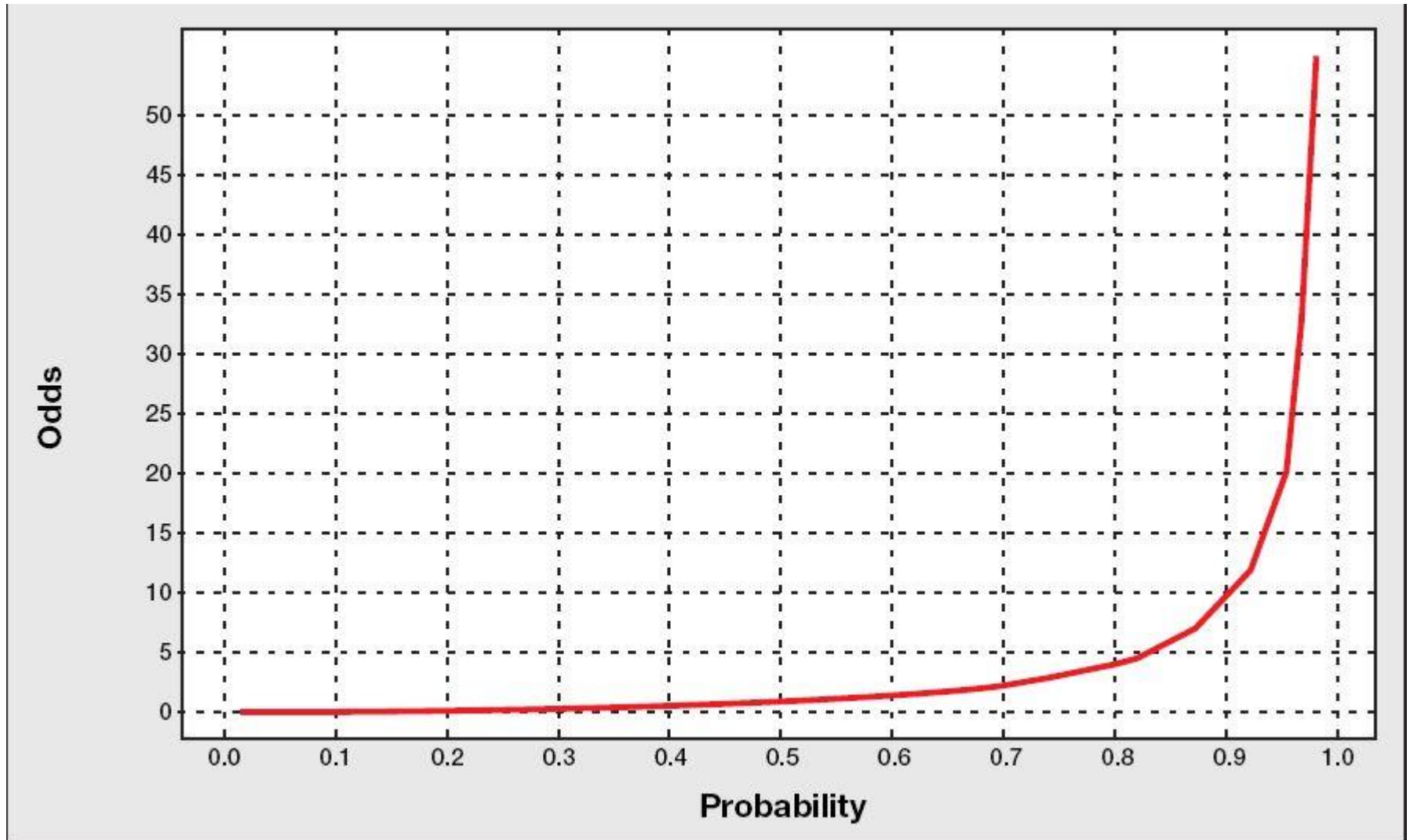
- p = probability of belonging to class 1 (success).

$$Odds = \frac{p}{1-p}$$

❖ 이전 예시에 대해

- 스페인의 우승 odds는 2/9이므로 스페인의 우승 확률은 2/11임
- 대한민국의 우승 odds는 1/250 이므로 대한민국의 우승확률은 $1/251 \approx 0.00398$ (0.398%)임
- 1,000년을 살면 대한민국이 월드컵에서 한 번 우승하는 모습을 목격할 수 있음

로지스틱 회귀분석: Odds



로지스틱 회귀분석: Log Odds

❖ Odds의 한계

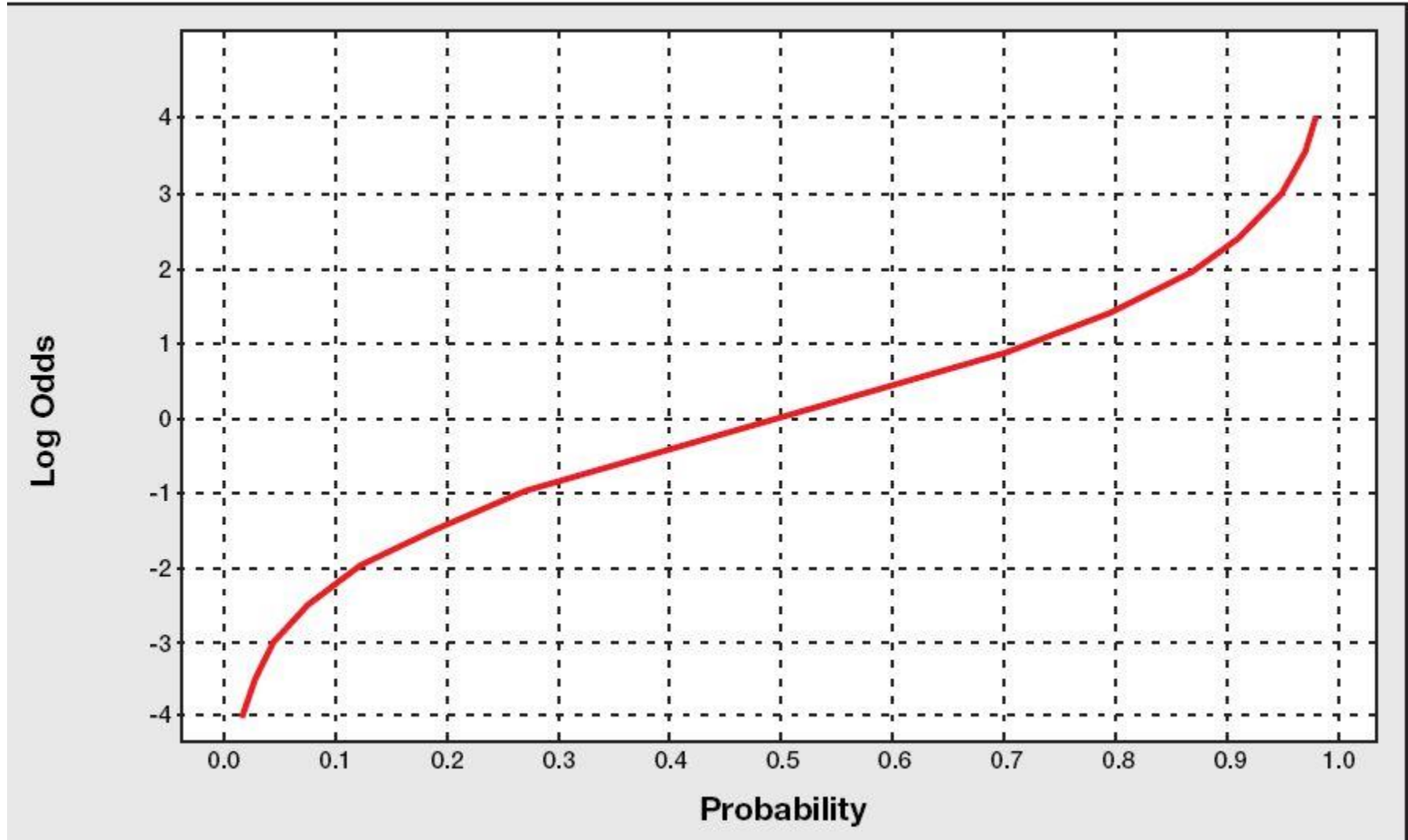
- 여전히 범위에 대한 제약이 존재함: $0 < \text{odds} < \infty$
- 비대칭성(Asymmetric)

❖ Odds에 로그를 취하자

$$\log(\text{Odds}) = \log\left(\frac{p}{1-p}\right)$$

- 드디어 범위에 대한 제약이 없어짐: $-\infty < \log(\text{odds}) < \infty$
- 대칭성 확보
- 성공확률 p 가 작으면 음수값을 갖고, 성공확률 p 가 크면 양수값을 가짐

로지스틱 회귀분석: Log Odds



로지스틱 회귀분석: Equation

❖ 로지스틱 회귀분석 식

- Log Odds를 이용한 회귀분석 식

$$\log(Odds) = \log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_n x_n$$

- 양변에 로그를 취하면

$$\frac{p}{1-p} = e^{\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_n x_n}$$

- 성공확률에 대한 식으로 표현

$$p = \sigma(\mathbf{x}, \boldsymbol{\beta}) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_n x_n)}}$$

로지스틱 회귀분석: Equation

❖ 로지스틱 회귀분석 식

Logistic
Regression
선형식

$$\ln\left(\frac{p}{1-p}\right) \quad : \text{logit} \quad (\text{odds에 자연로그를 취한 상태})$$

$$\ln\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x_1 + \cdots + \beta_i x_i$$

• 로지스틱 회귀 모형은 종속변수가 이분형일 때 선형회귀모형의 제약을 극복하기 위해 확률에 대한 로짓 변환을 고려하여 분석

$$P = \frac{\exp(\beta_0 + \beta_1 x_1 + \cdots + \beta_i x_i)}{1 + \exp(\beta_0 + \beta_1 x_1 + \cdots + \beta_i x_i)}$$

• 위의 모형식에서 추정된 회귀계수로부터 사후확률에 대한 추정식을 계산

로지스틱 회귀분석: 학습 (Optional)

❖ 최대 우도 추정법: Maximum likelihood estimation (MLE)

- Expectation function:

$$P(x_i, y_i | \beta) = \begin{cases} \sigma(x, \beta) & \text{if } y = 1 \\ 1 - \sigma(x, \beta) & \text{if } y = 0 \end{cases}$$

$$= \sigma(x, \beta)^y (1 - \sigma(x, \beta))^{1-y}$$

- Likelihood and log-likelihood of the training data \mathbf{X} :

$$L(\mathbf{X}, y, \beta) = \prod_{i=1}^R \sigma(x_i, \beta)^{y_i} (1 - \sigma(x_i, \beta))^{1-y_i}$$

$$\ln L(\mathbf{X}, y, \beta) = \sum_{i=1}^R y_i \ln(\sigma(x_i, \beta)) + (1 - y_i) \ln(1 - \sigma(x_i, \beta))$$

- 우도함수와 로그-우도함수는 회귀계수 β 에 대해 비선형이므로 선형회귀분석과 같이 명시적인 해가 존재하지 않음

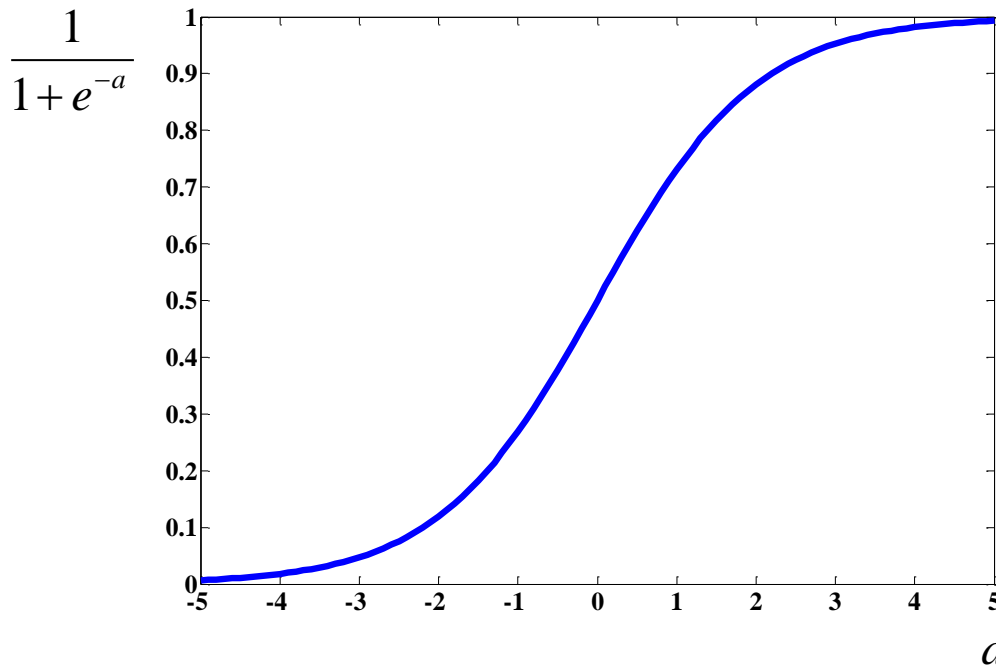
✓ Conjugate gradient 등의 최적화 알고리즘을 차용하여 해를 구함

로지스틱 회귀분석: 학습

❖ 성공 확률

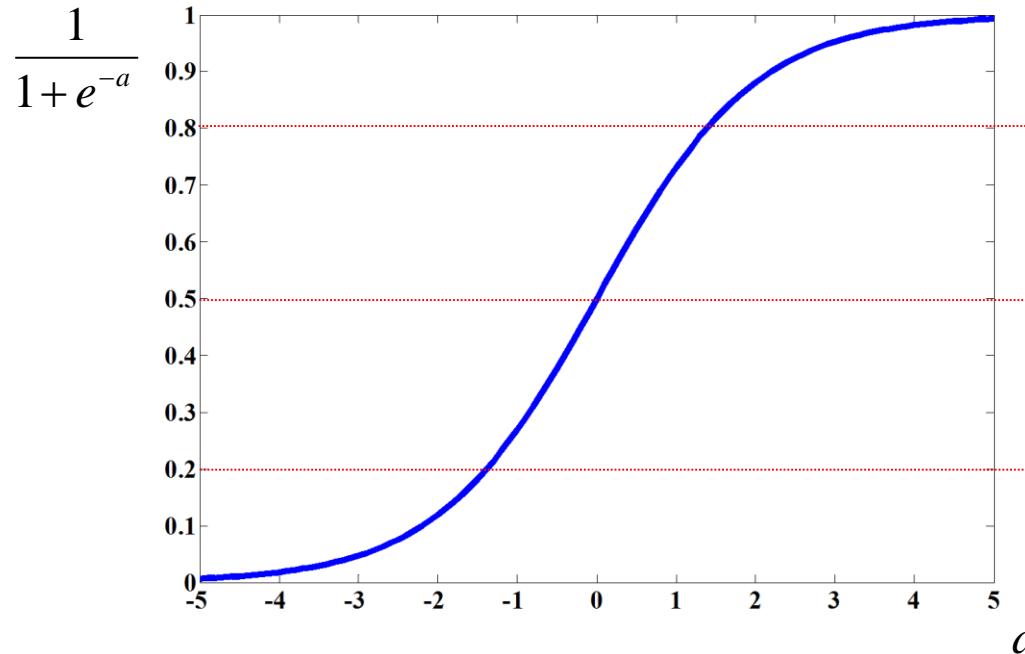
- 회귀계수가 추정되고 나면 주어진 설명변수집합에 대한 성공확률을 다음과 같이 계산할 수 있음

$$p = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_n x_n)}}$$



로지스틱 회귀분석: Cut-off

❖ 이진분류를 위한 cut-off 설정



성공 범주의 비중이 높을 때
(제조업 데이터에서는 거의
발생하지 않는 case)

가장 일반적인 cut-off

성공 범주의 비중이 낮을 때
(예: 불량 예측)

- 일반적으로 0.50가 주로 사용됨
- 사전확률을 고려한 cut-off나 검증데이터의 정확도를 최대화하는 cut-off 등이 사용될 수도 있음

로지스틱 회귀분석: Interpretation

❖ 로지스틱 회귀분석 회귀계수의 의미

■ 선형 회귀분석 회귀식

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \dots + \hat{\beta}_p x_p$$

✓ 선형 회귀분석에서의 회귀계수는 해당 변수가 1 증가함에 따른 **종속변수의 변화량**

■ 로지스틱 회귀분석 회귀식

$$\log(Odds) = \log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_p x_p$$

$$p = \sigma(\mathbf{x}, \beta) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_p x_p)}}$$

✓ 로지스틱 회귀분석에서의 회귀계수는 해당 변수가 1 증가함에 따른 **로그 승산의 변화량**

로지스틱 회귀분석: Interpretation

❖ 승산 비율: Odds Ratio

- 로지스틱 회귀분석에서 나머지 변수는 모두 고정시킨 상태에서 한 변수를 1만큼 증가시켰을 때 변화하는 Odds의 비율

- Odds ratio:

$$\frac{odds(x_1 + 1, \dots, x_n)}{odds(x_1, \dots, x_n)} = \frac{e^{\beta_0 + \beta_1(x_1+1) + \beta_2x_2 + \dots + \beta_nx_n}}{e^{\beta_0 + \beta_1x_1 + \beta_2x_2 + \dots + \beta_nx_n}} = e^{\beta_1}$$

- x_1 이 1 증가하게 되면 성공에 대한 승산 비율이 e^{β_1} 만큼 변화함
 - ✓ 회귀 계수가 양수 → 변수가 증가하면 성공 확률이 **증가** (성공범주와 **양의 상관관계**)
 - ✓ 회귀 계수가 음수 → 변수가 증가하면 성공 확률이 **감소** (성공범주와 **음의 상관관계**)

$$\frac{p}{1-p} = e^{\beta_0 + \beta_1x_1 + \beta_2x_2 + \dots + \beta_px_p}$$

로지스틱 회귀분석: Interpretation

❖ 로지스틱 회귀분석 결과 및 해석

- 로지스틱 회귀분석을 수행하고 나면 선형 회귀분석과 유사하게 다음과 같은 표를 결과로 얻을 수 있음

$$p = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_n x_n)}}$$

Input variables	Coefficient	Std. Error	p-value	Odds
Constant term	-13.20165825	2.46772742	0.00000009	*
Age	-0.04453737	0.09096102	0.62439483	0.95643985
Experience	0.05657264	0.09005365	0.5298661	1.05820346
Income	0.0657607	0.00422134	0	1.06797111
Family	0.57155931	0.10119002	0.00000002	1.77102649
CCAvg	0.18724874	0.06153848	0.00234395	1.20592725
Mortgage	0.00175308	0.00080375	0.02917421	1.00175464
Securities Account	-0.85484785	0.41863668	0.04115349	0.42534789
CD Account	3.46900773	0.44893095	0	32.10486984
Online	-0.84355801	0.22832377	0.00022026	0.43017724
CreditCard	-0.96406376	0.28254223	0.00064463	0.38134006
EducGrad	4.58909273	0.38708162	0	98.40509796
EducProf	4.52272701	0.38425466	0	92.08635712

로지스틱 회귀분석: Interpretation

❖ 로지스틱 회귀분석 결과 및 해석

■ 회귀계수: Coefficient

- ✓ 로지스틱 회귀분석에서 각 변수에 대응하는 베타값임
- ✓ 선형회귀분석에서는 해당 변수가 1단위 증가할 때 종속변수의 변화량을 의미하나, 로지스틱 회귀분석에서는 해당 변수가 1단위 증가할 때 로그승산비의 변화량을 의미
- ✓ 양수이면 성공확률과 양의 상관관계, 음수이면 성공 확률과 음의 상관관계

Input variables	Coefficient	Std. Error	p-value	Odds
Constant term	-13.20165825	2.46772742	0.00000009	*
Age	-0.04453737	0.09096102	0.62439483	0.95643985
Experience	0.05657264	0.09005365	0.5298661	1.05820346
Income	0.0657607	0.00422134	0	1.06797111
Family	0.57155931	0.10119002	0.00000002	1.77102649
CCAvg	0.18724874	0.06153848	0.00234395	1.20592725
Mortgage	0.00175308	0.00080375	0.02917421	1.00175464
Securities Account	-0.85484785	0.41863668	0.04115349	0.42534789
CD Account	3.46900773	0.44893095	0	32.10486984
Online	-0.84355801	0.22832377	0.00022026	0.43017724
CreditCard	-0.96406376	0.28254223	0.00064463	0.38134006
EducGrad	4.58909273	0.38708162	0	98.40509796
EducProf	4.52272701	0.38425466	0	92.08635712

로지스틱 회귀분석: Interpretation

❖ 로지스틱 회귀분석 결과 및 해석

■ 유의확률: p-value

- ✓ 로지스틱 회귀분석에서 해당 변수가 통계적으로 유의미한지 여부를 알려주는 지표
- ✓ 0에 가까울수록 모델링에 중요한 변수이며, 1에 가까울수록 유의미하지 않은 변수임
- ✓ 특정 유의수준(α)을 설정하여 해당 값 미만의 변수만을 사용하여 다시 로지스틱 회귀분석을 구축하는 것도 가능함 (주로 $\alpha = 0.05$ 사용)

Input variables	Coefficient	Std. Error	p-value	Odds
Constant term	-13.20165825	2.46772742	0.00000009	*
Age	-0.04453737	0.09096102	0.62439483	0.95643985
Experience	0.05657264	0.09005365	0.5298661	1.05820346
Income	0.0657607	0.00422134	0	1.06797111
Family	0.57155931	0.10119002	0.00000002	1.77102649
CCAvg	0.18724874	0.06153848	0.00234395	1.20592725
Mortgage	0.00175308	0.00080375	0.02917421	1.00175464
Securities Account	-0.85484785	0.41863668	0.04115349	0.42534789
CD Account	3.46900773	0.44893095	0	32.10486984
Online	-0.84355801	0.22832377	0.00022026	0.43017724
CreditCard	-0.96406376	0.28254223	0.00064463	0.38134006
EducGrad	4.58909273	0.38708162	0	98.40509796
EducProf	4.52272701	0.38425466	0	92.08635712

로지스틱 회귀분석: Interpretation

❖ 로지스틱 회귀분석 결과 및 해석

■ 승산 비율: Odds Ratio

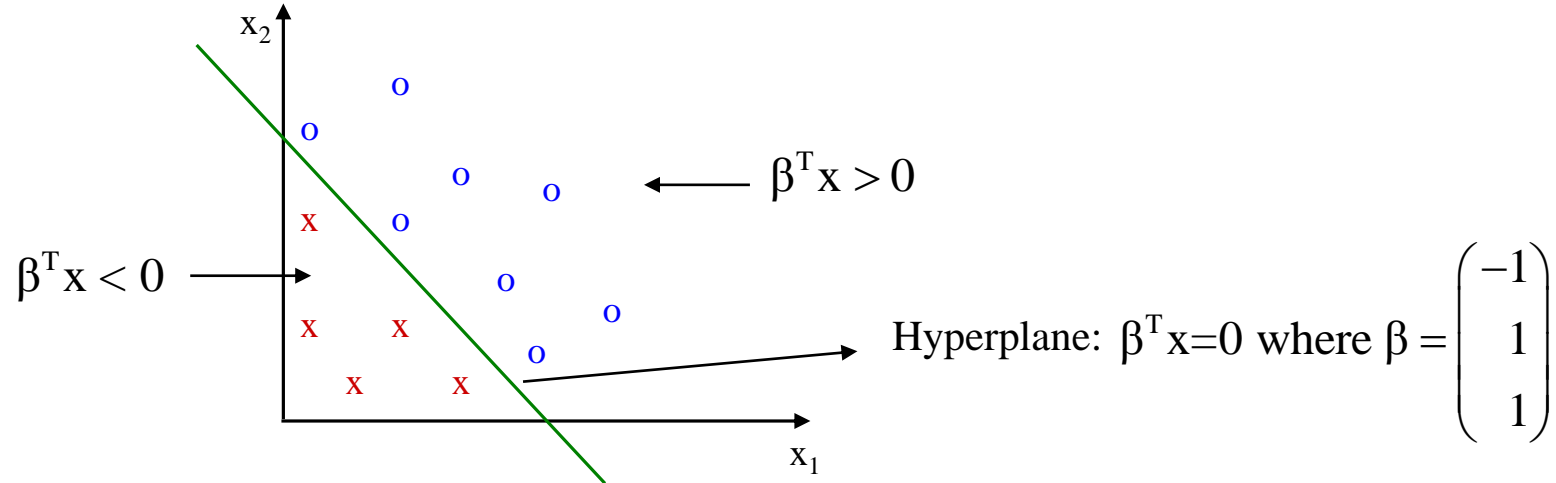
- ✓ 나머지 변수는 모두 고정시킨 상태에서 한 변수를 1만큼 증가시켰을 때 변화하는 Odds의 비율

Input variables	Coefficient	Std. Error	p-value	Odds
Constant term	-13.20165825	2.46772742	0.00000009	*
Age	-0.04453737	0.09096102	0.62439483	0.95643985
Experience	0.05657264	0.09005365	0.5298661	1.05820346
Income	0.0657607	0.00422134	0	1.06797111
Family	0.57155931	0.10119002	0.00000002	1.77102649
CCAvg	0.18724874	0.06153848	0.00234395	1.20592725
Mortgage	0.00175308	0.00080375	0.02917421	1.00175464
Securities Account	-0.85484785	0.41863668	0.04115349	0.42534789
CD Account	3.46900773	0.44893095	0	32.10486984
Online	-0.84355801	0.22832377	0.00022026	0.43017724
CreditCard	-0.96406376	0.28254223	0.00064463	0.38134006
EducGrad	4.58909273	0.38708162	0	98.40509796
EducProf	4.52272701	0.38425466	0	92.08635712

로지스틱 회귀분석: Interpretation

❖ Geometric interpretation

- 로지스틱 회귀분석은 d 차원의 데이터를 구분하는 $(d-1)$ 차원의 초평면을 찾는 것으로 이해할 수 있음



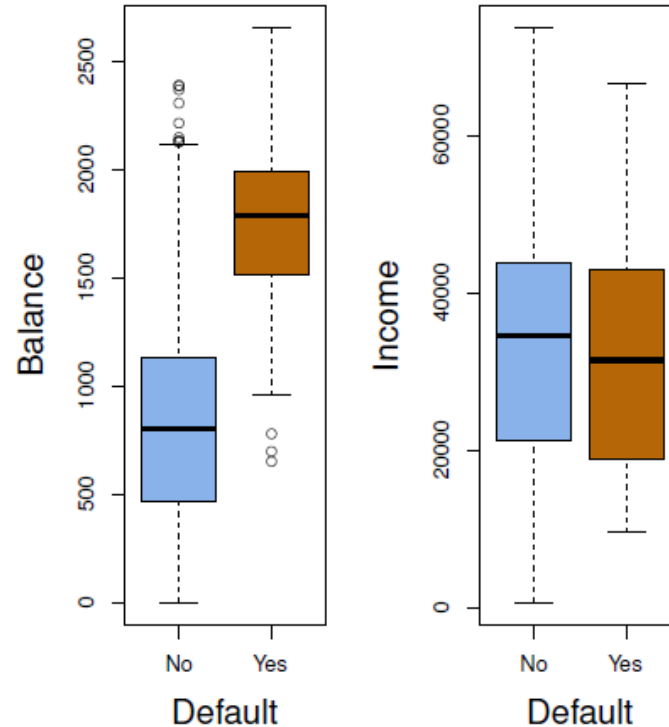
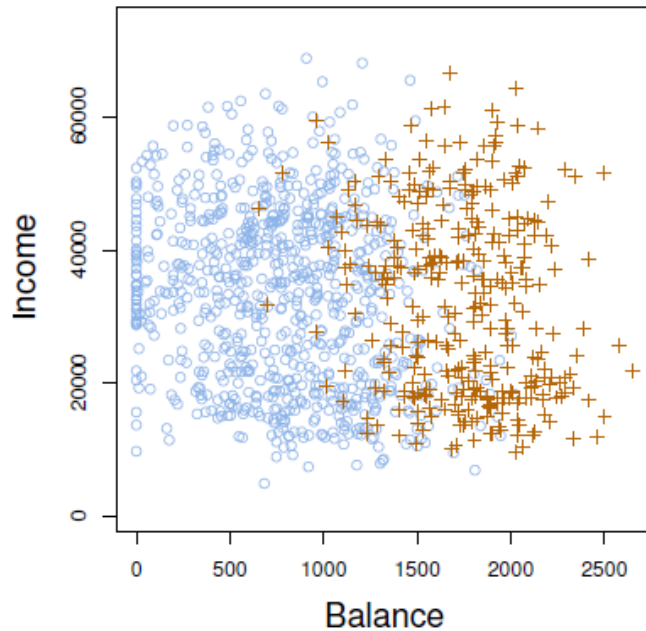
Classifier

$$y = \frac{1}{(1 + \exp(-\beta^T x))}$$

$$\begin{cases} y \rightarrow 1 & \text{if } \beta^T x \rightarrow \infty \\ y = \frac{1}{2} & \text{if } \beta^T x = 0 \\ y \rightarrow 0 & \text{if } \beta^T x \rightarrow -\infty \end{cases}$$

로지스틱 회귀분석: 학습

❖ 신용카드 연체 예측



$$p(X) = \frac{e^{\beta_0 + \beta_1 X}}{1 + e^{\beta_0 + \beta_1 X}}.$$

$$\log \left(\frac{p(X)}{1 - p(X)} \right) = \beta_0 + \beta_1 X.$$

로지스틱 회귀분석: 학습

❖ Credit Card Default: single variable

	Coefficient	Std. Error	Z-statistic	P-value
Intercept	-10.6513	0.3612	-29.5	< 0.0001
balance	0.0055	0.0002	24.9	< 0.0001

What is our estimated probability of **default** for someone with a balance of \$1000?

$$\hat{p}(X) = \frac{e^{\hat{\beta}_0 + \hat{\beta}_1 X}}{1 + e^{\hat{\beta}_0 + \hat{\beta}_1 X}} = \frac{e^{-10.6513 + 0.0055 \times 1000}}{1 + e^{-10.6513 + 0.0055 \times 1000}} = 0.006$$

With a balance of \$2000?

$$\hat{p}(X) = \frac{e^{\hat{\beta}_0 + \hat{\beta}_1 X}}{1 + e^{\hat{\beta}_0 + \hat{\beta}_1 X}} = \frac{e^{-10.6513 + 0.0055 \times 2000}}{1 + e^{-10.6513 + 0.0055 \times 2000}} = 0.586$$

로지스틱 회귀분석: 학습

❖ Credit Card Default: multiple variables

$$\log \left(\frac{p(X)}{1 - p(X)} \right) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p$$

$$p(X) = \frac{e^{\beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p}}{1 + e^{\beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p}}$$

	Coefficient	Std. Error	Z-statistic	P-value
Intercept	-10.8690	0.4923	-22.08	< 0.0001
balance	0.0057	0.0002	24.74	< 0.0001
income	0.0030	0.0082	0.37	0.7115
student [Yes]	-0.6468	0.2362	-2.74	0.0062

목차

I

나이프 베이즈 분류기

II

로지스틱 회귀분석

III

로지스틱 회귀분석 활용사례: 마케팅 반응 모델링

IV

R 실습

로지스틱 회귀분석: 예시

❖ 개인 신용대출 예측

- 개인의 인구통계학적인 정보와 은행이용행태를 바탕으로 신용대출 여부를 판별

일련 번호	나이	경력	소득	가족 수	월별 신용카드 평균사용액	교육 수준	담보부 채권	개인 대출	증권 계좌	CD 계좌	온라인 뱅킹	신용 카드
1	25	1	49	4	1.60	UG	0	No	Yes	No	No	No
2	45	19	34	3	1.50	UG	0	No	Yes	No	No	No
3	39	15	11	1	1.00	UG	0	No	No	No	No	No
4	35	9	100	1	2.70	Grad	0	No	No	No	No	No
5	35	8	45	4	1.00	Grad	0	No	No	No	No	Yes
6	37	13	29	4	0.40	Grad	155	No	No	No	Yes	No
7	53	27	72	2	1.50	Grad	0	No	No	No	Yes	No
8	50	24	22	1	0.30	Prof	0	No	No	No	No	Yes
9	35	10	81	3	0.60	Grad	104	No	No	No	Yes	No
10	34	9	180	1	8.90	Prof	0	Yes	No	No	No	No
11	65	39	105	4	2.40	Prof	0	No	No	No	No	No
12	29	5	45	3	0.10	Grad	0	No	No	No	Yes	No
13	48	23	114	2	3.80	Prof	0	No	Yes	No	No	No
14	59	32	40	4	2.50	Grad	0	No	No	No	Yes	No
15	67	41	112	1	2.00	UG	0	No	Yes	No	No	No
16	60	30	22	1	1.50	Prof	0	No	No	No	Yes	Yes
17	38	14	130	4	4.70	Prof	134	Yes	No	No	No	No
18	42	18	81	4	2.40	UG	0	No	No	No	No	No
19	46	21	193	2	8.10	Prof	0	Yes	No	No	No	No
20	55	28	21	1	0.50	Grad	0	No	Yes	No	No	Yes

로지스틱 회귀분석: 예시

I

데이터 전처리

- 총 5,000명의 고객 대상
 - 설명변수
 - ✓ 인구통계학적 변수: age, income, etc.
 - ✓ 은행이용행태 변수: mortgage, security account, etc.
 - 480(9.6%)의 고객만이 신용대출을 받음
-
- 60% for training, 40% for validation.
 - Create dummy variables for the categorical predictors.
$$EducProf = \begin{cases} 1 & \text{if education is } Professional \\ 0 & \text{otherwise} \end{cases}$$
$$EducGrad = \begin{cases} 1 & \text{if education is at } Graduate \text{ level} \\ 0 & \text{otherwise} \end{cases}$$

로지스틱 회귀분석: 예시

2

단일 변수를 이용한 로지스틱 회귀분석

Modeling with “Income” only

$$\text{Prob}(\text{Personal Loan} = \text{Yes} \mid \text{Income} = x) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x)}}$$

$$\text{Odds}(\text{Personal Loan} = \text{Yes}) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x)}}$$

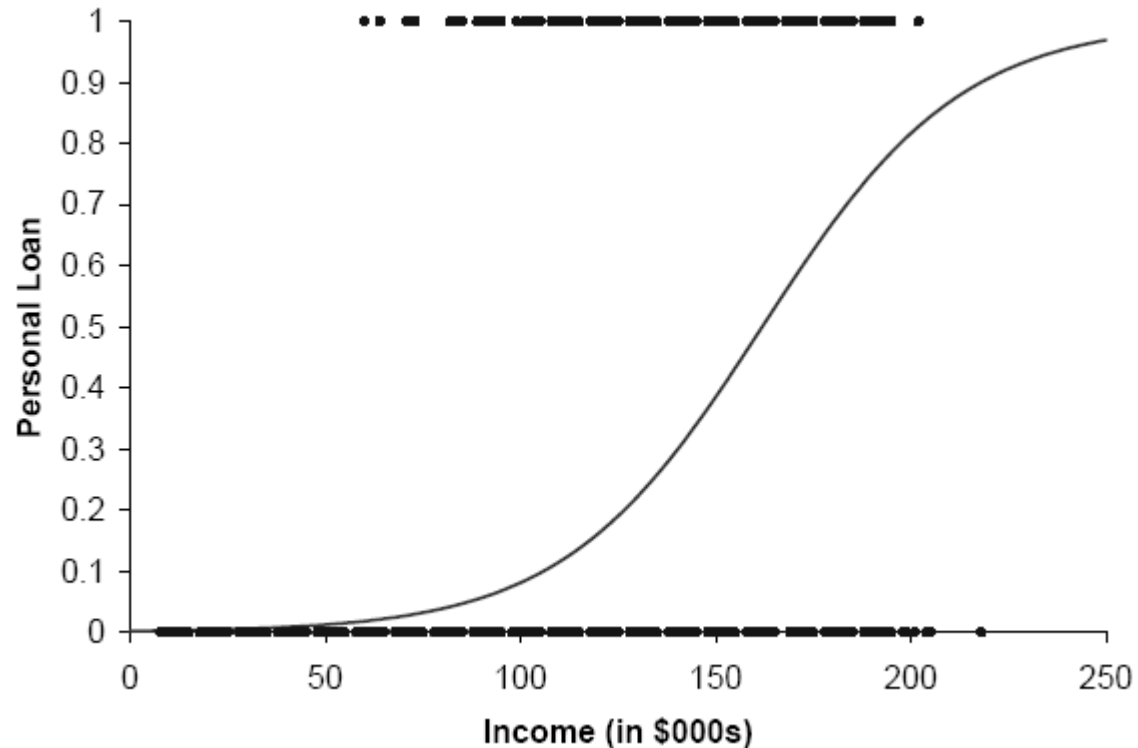
$$P(\text{Personal Loan} = \text{Yes} \mid \text{Income} = x) = \frac{1}{1 + e^{-(-6.3525 + 0.0392x)}}$$

로지스틱 회귀분석: 예시

2

단일 변수를 이용한 로지스틱 회귀분석

Modeling with “Income” only



로지스틱 회귀분석: 예시

2

모든 변수를 이용한 로지스틱 회귀분석

$$p = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_n x_n)}}$$

Input variables	Coefficient	Std. Error	p-value	Odds
Constant term	-13.20165825	2.46772742	0.00000009	*
Age	-0.04453737	0.09096102	0.62439483	0.95643985
Experience	0.05657264	0.09005365	0.5298661	1.05820346
Income	0.0657607	0.00422134	0	1.06797111
Family	0.57155931	0.10119002	0.00000002	1.77102649
CCAvg	0.18724874	0.06153848	0.00234395	1.20592725
Mortgage	0.00175308	0.00080375	0.02917421	1.00175464
Securities Account	-0.85484785	0.41863668	0.04115349	0.42534789
CD Account	3.46900773	0.44893095	0	32.10486984
Online	-0.84355801	0.22832377	0.00022026	0.43017724
CreditCard	-0.96406376	0.28254223	0.00064463	0.38134006
EducGrad	4.58909273	0.38708162	0	98.40509796
EducProf	4.52272701	0.38425466	0	92.08635712

로지스틱 회귀분석: 예시

예측 수행

$$p = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_n x_n)}}$$

3

승률계산을 위한 분류기준값(갱신가능)		0.5	(여기에 있는 분류기준값을 변경할지라도 요약보고서의 분류기준값은 바뀌지 않는다)					
행번호	예측집단	실제집단	성공확률	로그 승산	나 이	경 력	소 득	가족 수
2	0	0	2.1351E-05	-10.75439275	45	19	34	3
3	0	0	3.34564E-06	-12.60785033	39	15	11	1
7	0	0	0.015822384	-4.13038073	53	27	72	2
8	0	0	0.000216511	-8.437650808	50	24	22	1
11	1	0	0.567824439	0.272980386	65	39	105	4

목차

I

나이프 베이즈 분류기

II

로지스틱 회귀분석

III

로지스틱 회귀분석 활용사례: 마케팅 반응 모델링

IV

R 실습

Naïve Bayes & Logistic Regression: R Exercise

❖ Personal Loan Prediction

Data Description:

ID	Customer ID
Age	Customer's Age in completed years
Experience	#years of professional experience
Income	Annual income of the customer (\$000)
ZIPCode	Home Address ZIP code.
Family	Family size (dependents) of the customer
CCAvg	Avg. Spending on Credit Cards per month (\$000)
Education	Education Level. 1: Undergrad; 2: Graduate; 3: Advanced/Professional
Mortgage	Value of house mortgage if any. (\$000)
Personal Loan	Did this customer accept the personal loan offered in the last campaign?
Securities Account	Does the customer have a Securities account with the bank?
CD Account	Does the customer have a Certificate of Deposit (CD) account with the bank?
Online	Does the customer use internet banking facilities?
CreditCard	Does the customer use a credit card issued by UniversalBank?

Naïve Bayesian Classification

❖ package 'e1071'

Package 'e1071'

March 17, 2014

Version 1.6-3

Date 2014-02-13

Title Misc Functions of the Department of Statistics (e1071), TU Wien

Imports graphics, grDevices, class

Suggests cluster, mlbench, nnet, randomForest, rpart, SparseM, xtable, Matrix, MASS

Description Functions for latent class analysis, short time Fourier transform, fuzzy clustering, support vector machines, shortest path computation, bagged clustering, naive Bayes classifier, ...

License GPL-2

LazyLoad yes

Author David Meyer [aut, cre], Evgenia Dimitriadou [aut, cph], Kurt Hornik [aut], Andreas Weingessel [aut], Friedrich Leisch [aut], Chih-Chung Chang [ctb, cph] (libsvm C++-code), Chih-Chen Lin [ctb, cph] (libsvm C++-code)

Maintainer David Meyer <David.Meyer@R-project.org>

NeedsCompilation yes

Repository CRAN

Date/Publication 2014-03-17 12:12:07

Naïve Bayesian Classification

❖ 패키지 로드 및 데이터 준비

- Zip code, idx 등 필요 없는 입력 변수 제거

```

17 # Naive Bayesian Classifier -----
18 # e1071 package install
19 install.packages("e1071", dependencies = TRUE)
20
21 # call the e1071 package
22 library(e1071)
23
24 ploan <- read.csv("Personal Loan.csv")
25
26 input_idx <- c(2,3,4,6,7,8,9,11,12,13,14)
27 target_idx <- 10
28
29 ploan_input <- ploan[,input_idx]
30 ploan_target <- as.factor(ploan[,target_idx])
31 ploan_data <- data.frame(ploan_input, ploan_target)
32
33 # split the data into the training/validation sets
34 trn_idx <- sample(1:dim(ploan_data)[1], round(0.7*dim(ploan_data)[1]))
35 ploan_trn <- ploan_data[trn_idx,]
36 ploan_val <- ploan_data[-trn_idx,]

```

Naïve Bayesian Classification

❖ 나이트 베이즈 분류기 학습

```

38 # Training the Naive Bayesian Classifier
39 nb_model <- naiveBayes(ploan_target ~ ., data = ploan_trn)
40 nb_model$apriori
41 nb_model$tables

```

```
> nb_model$apriori
```

```
Y
  0    1
1564 186
```

→ prior distribution

```
> nb_model$tables
```

```
$Age
  Age
Y    [,1]    [,2]
0 45.43286 11.62866
1 45.85484 11.34974
```

```
$Experience
  Experience
Y    [,1]    [,2]
0 20.20013 11.63852
1 20.67742 11.46620
```

→ [1]: mean
[2]: standard deviation

```
$Income
  Income
Y    [,1]    [,2]
0 66.67647 41.55169
1 146.00538 32.73258
```

Naïve Bayesian Classification

❖ 분류 성능 평가

```

43 # Predict the new input data based on Naive Bayesian Classifier
44 posterior = predict(nb_model, ploan_val, type = "raw")
45 nb_prej = predict(nb_model, ploan_val, type = "class")
46
47 # Generate a confusion matrix
48 cfmatrix <- table(ploan_val$ploan_target, nb_prej)
49
50 # Evaluate the performance
51 perf_mat <- matrix(0, 4, 3)
52
53 perf_mat[,1] <- perf_eval(cfmatrix)

```

■ TPR, TNR, ACC, BCR 계산 함수

```

1 # Performance Evaluation Function -----
2 perf_eval <- function(cm){
3
4   # True positive rate: TPR
5   TPR = cm[2,2]/sum(cm[2,])
6   # True negative rate: TNR
7   TNR = cm[1,1]/sum(cm[1,])
8   # Simple Accuracy
9   ACC = (cm[1,1]+cm[2,2])/sum(cm)
10  # Balanced Correction Rate
11  BCR = sqrt(TPR*TNR)
12
13  return(c(TPR, TNR, ACC, BCR))
14 }

```

```

> perf_mat
      [,1] [,2] [,3]
[1,] 0.5142857 0 0
[2,] 0.9191176 0 0
[3,] 0.8813333 0 0
[4,] 0.6875239 0 0

```

Logistic Regression: 데이터 정규화

❖ 데이터 정규화

```

55 # Logistic Regression -----
56 # Conduct the normalization
57 ploan_input <- scale(ploan_input, center = TRUE, scale = TRUE)
58 ploan_target <- as.numeric(ploan_target)-1
59 ploan_data <- data.frame(ploan_input, ploan_target)
60
61 ploan_trn <- ploan_data[trn_idx,]
62 ploan_val <- ploan_data[-trn_idx,]

```

Age	Experience	Income	Family	CCAvg	Education	Mortgage	Securities.Account	CD.Account	Online	CreditCard	ploan_target
1.18529234	1.20439184	-0.69521173	0.5104066	-0.25134880	-1.0347255	-0.5692367	2.8497084	3.8755168	-1.2183987	1.5628656	0
0.23039152	0.24984761	-1.38084178	-0.3517667	-0.69691974	1.3560450	-0.5692367	-0.3507727	-0.2579269	0.8204211	1.5628656	0
-1.15855512	-1.05180361	1.42595623	-0.3517667	1.32485842	-1.0347255	-0.5692367	-0.3507727	-0.2579269	-1.2183987	-0.6395944	0
0.49081902	0.59695461	-1.10230457	-1.2139401	-0.97540158	-1.0347255	-0.5692367	-0.3507727	-0.2579269	0.8204211	-0.6395944	0
-1.24536428	-1.31213385	1.08314121	-0.3517667	-0.36274153	-1.0347255	-0.5692367	-0.3507727	-0.2579269	0.8204211	1.5628656	0
0.40400985	0.51017786	-1.12373051	1.3725800	-0.80831248	0.1606598	0.5416936	-0.3507727	-0.2579269	0.8204211	-0.6395944	0
1.01167401	1.03083835	-0.65235985	-0.3517667	-0.19565243	1.3560450	-0.5692367	-0.3507727	-0.2579269	-1.2183987	-0.6395944	0
-1.50579178	-1.48568735	0.33323334	-0.3517667	0.75118583	-1.0347255	-0.5692367	-0.3507727	-0.2579269	0.8204211	1.5628656	0
-0.29046347	-0.27081288	-0.90947112	1.3725800	-1.08679432	0.1606598	-0.5692367	-0.3507727	-0.2579269	0.8204211	-0.6395944	0
1.44571984	1.46472209	-0.52380422	0.5104066	0.24991852	-1.0347255	-0.5692367	-0.3507727	-0.2579269	0.8204211	-0.6395944	0

Logistic Regression: 데이터 정규화

❖ Data preparation

	Age	Experience	Income	Family	CCAvg	Education	Mortgage	Securities.Account	CD.Account	Online	CreditCard	Ploan
1	-1.76621927	-1.65924085	-0.54523016	1.3725800	-0.19565243	-1.0347255	-0.5692367	1	0	0	0	0
2	-0.03003597	-0.09725938	-0.86661924	0.5104066	-0.25134880	-1.0347255	-0.5692367	1	0	0	0	0
3	-0.55089096	-0.44436637	-1.35941584	-1.2139401	-0.52983064	-1.0347255	-0.5692367	0	0	0	0	0
4	-0.89812762	-0.96502686	0.54749273	-1.2139401	0.41700762	0.1606598	-0.5692367	0	0	0	0	0
5	-0.89812762	-1.05180361	-0.63093391	1.3725800	-0.52983064	0.1606598	-0.5692367	0	0	0	1	0
6	-0.72450929	-0.61791987	-0.97374894	1.3725800	-0.86400885	0.1606598	0.9682115	0	0	1	0	0
7	0.66443735	0.59695461	-0.05243356	-0.3517667	-0.25134880	0.1606598	-0.5692367	0	0	1	0	0
8	0.40400985	0.33662436	-1.12373051	-1.2139401	-0.91970522	1.3560450	-0.5692367	0	0	0	1	0
9	-0.89812762	-0.87825011	0.14039989	0.5104066	-0.75261611	0.1606598	0.4623414	0	0	1	0	0
10	-0.98493679	-0.96502686	2.26156786	-1.2139401	3.87018245	1.3560450	-0.5692367	0	0	0	0	1
11	1.70614733	1.63827558	0.65462243	1.3725800	0.24991852	1.3560450	-0.5692367	0	0	0	0	0
12	-1.41898261	-1.31213385	-0.63093391	0.5104066	-1.03109795	0.1606598	-0.5692367	0	0	1	0	0
13	0.23039152	0.24984761	0.84745588	-0.3517667	1.02966767	1.3560450	-0.5692367	1	0	0	0	0
14	1.18529234	1.03083835	-0.73806361	1.3725800	0.30561488	0.1606598	-0.5692367	0	0	1	0	0
15	1.87976566	1.81182908	0.80460400	-1.2139401	0.02713304	-1.0347255	-0.5692367	1	0	0	0	0
16	1.27210151	0.85728485	-1.12373051	-1.2139401	-0.25134880	1.3560450	-0.5692367	0	0	1	1	0
17	-0.63770013	-0.53114312	1.19027091	1.3725800	1.53093498	1.3560450	0.7599121	0	0	0	0	1
18	-0.29046347	-0.18403613	0.14039989	1.3725800	0.24991852	-1.0347255	-0.5692367	0	0	0	0	0
19	0.05677319	0.07629412	2.54010506	-0.3517667	3.42461150	1.3560450	-0.5692367	0	0	0	0	1
20	0.83805568	0.68373135	-1.14515645	-1.2139401	-0.80831248	0.1606598	-0.5692367	1	0	0	1	0
21	0.92486485	0.94406160	-1.05945269	1.3725800	-0.58552701	0.1606598	0.5317746	0	0	1	0	0
22	1.01167401	0.59695461	-0.24526701	0.5104066	0.02713304	1.3560450	-0.5692367	0	0	1	0	0
23	-1.41898261	-1.31213385	-0.26669295	-1.2139401	-0.41843790	-1.0347255	2.0097087	0	0	1	0	0
24	-0.11684514	-0.18403613	-0.67378579	-0.3517667	-0.69691974	-1.0347255	1.0475637	1	0	0	0	0
25	-0.81131846	-0.79147336	1.66164156	-0.3517667	1.08536404	-1.0347255	1.0078876	0	0	0	1	0

Logistic Regression: 모든 변수를 이용하여 학습

❖ 모든 변수를 이용한 로지스틱 회귀분석

■ Generalized linear model (glm)을 이용하여 로지스틱 회귀분석 실시

```
64 # Train the Logistic Regression Model with all variables
65 full_lr <- glm(plan_target ~ ., family=binomial, plan_trn)
66 full_lr
67 summary(full_lr)
> summary(full_lr)
```

```
Call:
glm(formula = plan_target ~ ., family = binomial, data = plan_trn)
```

Deviance Residuals:

Min	1Q	Median	3Q	Max
-2.2192	-0.2173	-0.0870	-0.0349	3.8473

Coefficients:

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	-4.45619	0.25571	-17.427	< 2e-16 ***
Age	0.73706	1.08882	0.677	0.498443
Experience	-0.44795	1.08252	-0.414	0.679020
Income	2.47260	0.19753	12.518	< 2e-16 ***
Family	0.84143	0.14377	5.853	4.84e-09 ***
CCAvg	0.21323	0.11498	1.854	0.063670 .
Education	1.35708	0.15736	8.624	< 2e-16 ***
Mortgage	-0.02584	0.09805	-0.264	0.792104
Securities.Account	-0.42369	0.15448	-2.743	0.006095 **
CD.Account	0.84216	0.13111	6.423	1.33e-10 ***
Online	-0.15084	0.12779	-1.180	0.237826
Creditcard	-0.54431	0.15999	-3.402	0.000669 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 1150.92 on 1749 degrees of freedom
 Residual deviance: 467.52 on 1738 degrees of freedom
 AIC: 491.52

Number of Fisher Scoring iterations: 7

Logistic Regression: 후방소거법을 이용한 변수선택

❖ 후방소거법을 이용한 변수 선택

```
69 # Train the Logistic Regression Model with selected variables
70 reduced_lr <- stepAIC(full_lr, direction = "backward")
71 summary(reduced_lr)
```

```
> summary(reduced_lr)
```

```
Call:
glm(formula = ploan_target ~ Age + Income + Family + CCAvg +
    Education + Securities.Account + CD.Account + CreditCard,
    family = binomial, data = ploan_trn)
```

Deviance Residuals:

Min	1Q	Median	3Q	Max
-2.2101	-0.2197	-0.0896	-0.0357	3.8096

Coefficients:

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	-4.4335	0.2531	-17.518	< 2e-16 ***
Age	0.2905	0.1274	2.281	0.022565 *
Income	2.4639	0.1949	12.642	< 2e-16 ***
Family	0.8441	0.1423	5.931	3.01e-09 ***
CCAvg	0.2170	0.1141	1.902	0.057140 .
Education	1.3660	0.1531	8.923	< 2e-16 ***
Securities.Account	-0.4118	0.1527	-2.698	0.006981 **
CD.Account	0.7990	0.1250	6.391	1.65e-10 ***
CreditCard	-0.5245	0.1575	-3.331	0.000865 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 1150.9 on 1749 degrees of freedom
Residual deviance: 469.2 on 1741 degrees of freedom
AIC: 487.2

Number of Fisher scoring iterations: 7

Logistic Regression: 예측 성능 비교

❖ 모델 성능 평가

■ 검증 데이터에 대한 성능 평가 지표 계산

```

73 # Evaluate the logistic regression performance on the validation data
74 # Case 1: full model
75 full_response <- predict(full_lr, type = "response", newdata = ploan_val)
76 full_target <- ploan_val$plloan_target
77 full_predicted <- rep(0, length(full_target))
78 full_predicted[which(full_response >= 0.5)] <- 1
79 cm_full <- table(full_target, full_predicted)
80
81 perf_mat[,2] <- perf_eval(cm_full)
82
83 # Case 2: reduced model
84 reduced_response <- predict(reduced_lr, type = "response", newdata = ploan_val)
85 reduced_target <- ploan_val$plloan_target
86 reduced_predicted <- rep(0, length(reduced_target))
87 reduced_predicted[which(reduced_response >= 0.5)] <- 1
88 cm_reduced <- table(reduced_target, reduced_predicted)
89
90 perf_mat[,3] <- perf_eval(cm_reduced)
91 colnames(perf_mat) <- c("Naive Bayes", "LR with all variables", "LR with selected variables")
92 rownames(perf_mat) <- c("TPR", "TNR", "ACC", "BCR")
93 perf_mat

```

```

> perf_mat
      Naive Bayes LR with all variables LR with selected variables
TPR    0.6025641          0.6282051          0.6282051
TNR    0.9092262          0.9866071          0.9880952
ACC    0.8773333          0.9493333          0.9506667
BCR    0.7401804          0.7872685          0.7878620

```

Q & A

